

Tilburg University

Dynamische componenten analyse

Stobberingh, Robert

Publication date:
1972

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Stobberingh, R. (1972). *Dynamische componenten analyse: een integratie van componenten- en tijdreeksanalyse*. [s.n.].

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

CBM

71

F

101A

DYNAMISCHE COMPONENTEN ANALYSE

EEN INTEGRATIE VAN COMPONENTEN- EN TIJDREEKSANALYSE

R. STOBBERINGH

STELLINGEN

I

Om nieuwe inzichten in het economisch gebeuren te introduceren in regressiemodellen, dienen nieuwe variabelen in deze modellen te worden ingevoerd, welke deze nieuwe inzichten weerspiegelen. Om de samenhang tussen de traditionele en de nieuwe variabelen op te sporen verdient het aanbeveling, alle variabelen vooraf aan een componenten- of factoranalyse te onderwerpen, teneinde meer zekerheid te krijgen over de aard en het aantal van de in de modellen te hanteren structurele relaties.

II

Met het oog op hetgeen in stelling I naar voren is gebracht, verdient het aanbeveling het onderwijs in de statistiek aan de Economische Faculteiten uit te breiden met multivariate analyse methoden, zoals bijvoorbeeld de componenten- of factoranalyse.

III

De aanwezigheid van een variabele met een hoge specifieke variantie component, kan een aanwijzing zijn voor het feit, dat deze variabele samenhang vertoont met variabelen, welke nog niet in het onderzoek betrokken zijn.

Dat wil dus zeggen, dat de betreffende variabele niet thuisheert in het model, òf dat het model uitgebreid dient te worden met andere, nieuwe variabelen.

IV

Bij de in een model geformaliseerde theorie komt vaak te eenzijdig de nadruk te liggen op de oplossingsmethode, terwijl aan de verificatie van de aan het model ten grondslag liggende vooronderstellingen en de randvoorwaarden welke het model conditioneren, nauwelijks aandacht wordt besteed. Hierdoor ontstaat het gevaar dat het model zijn betekenis voor het probleem, waarvoor het is opgesteld, verliest.

V

Het complexe karakter van economische macro- en micro systemen, en de grote mate van onderlinge afhankelijkheid welke tussen deze systemen bestaat, maakt dat "toepassing van de cybernetica in het economisch onderzoek" eerder een poging inhoudt de cybernetische denkwijze te introduceren met een hierop geöriënteerde methode van onderzoek, dan dat er van een werkelijk toepassen sprake kan zijn.

VI

De bijdrage van de spectraalanalyse aan de econometrie is slechts van zeer beperkte betekenis.

VII

Een belangrijk probleem in de optimaliseringstheorie is het elimineren van de ruis welke op een signaal is gesuperponeerd, om zodoende het oorspronkelijke signaal te verkrijgen. Dit houdt een transformatie in van het waargenomen, door ruis verstoorte signaal, met behulp van een gewichtsfunctie $g(s,t)$ welke moet worden opgelost uit de volgende integraalvergelijking:

$$\int_T K_x(t,\tau) g(s,t) dt = f(s,\tau)$$

Hierin is $K_x(t,\tau)$ de correlatiefunctie van de ruis $\underline{x}(t)$. De bepaling van $g(s,t)$ uit bovenstaande integraalvergelijking zou bijzonder eenvoudig geschieden indien $\underline{x}(t)$ witte ruis zou zijn. Dit betekent namelijk dat:

$$K_x(t,\tau) = G(t) \delta(t-\tau),$$

waarin $G(t)$ de intensiteit van witte ruis voorstelt. De oplossing is:

$$g(s,\tau) = \frac{f(s,\tau)}{G(\tau)}.$$

Deze gedachtegang geeft aanleiding tot de volgende methode ter bepaling van de gewichtsfunctie:

1. Bepaal die functie die het stochastisch proces $\underline{x}(t)$ transformeert tot witte ruis.
Dit geschiedt met behulp van de methode van de canonieke ontwikkeling van stochastische processen.
2. Bereken de optimale gewichtsfunctie met witte ruis als ingangssignaal.
3. De optimale gewichtsfunctie $g(s,t)$ met $\underline{x}(t)$ als ingangssignaal wordt dan verkregen door serieschakeling van de in ad. 1 en ad. 2 bepaalde systemen.

VIII

Interpolatie en extrapolatie van tijdreeksen kan succesvol geschieden, met behulp van de in de vorige stelling beschreven procedure.

IX

Regelmatige wijzigingen en aanpassingen in het studieprogramma, wekken de indruk dat er sprake zou zijn van een naar een optimum convergerend proces.

Het is echter pijnlijk te moeten vaststellen, dat men niet weet waar dit optimum ligt, zodat men ook nooit zal weten of dit optimum bereikt is.

Stellingen behorende bij het proefschrift van R. Stobberingh:
Dynamische Componenten Analyse.

BIBLIOTHEEK KATHOLIEKE HOGESCHOOL
Hogeschoollaan 225, Tilburg - Tel. 04250-69111

Dit werk terug te bezorgen uiterlijk op :

BEPALING UIT HET REGLEMENT

Een werk, dat iemand in bruikleen heeft, mag
door hem in geen geval worden uitgeleend.

DYNAMISCHE COMPONENTEN ANALYSE

Een integratie van componenten- en tijdreeksanalyse

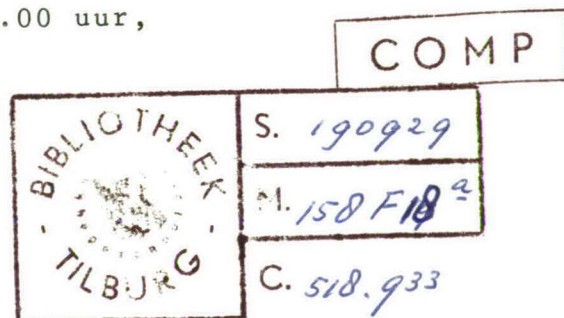
PROEFSCHRIFT

ter verkrijging van de graad van doctor in de
economische wetenschappen aan de Katholieke
Hogeschool te Tilburg, op gezag van de Rector
Magnificus Prof.Dr.C.F.Scheffer, in het open-
baar te verdedigen ten overstaan van een door
het College van Decanen aangewezen Commissie
in de aula van de Hogeschool op woensdag
24 mei 1972, des namiddags te 16.00 uur,

door

Robert Stobberingh

geboren te Batavia



DYNAMISCHE COMPONENTEN ANALYSE

Een integratie van componenten- en tijdreeksanalyse

R.Stobberingh

Promotor: Prof.Dr.J.J.J.Dalmulder

Aan mijn ouders

The mind may, as it appears to me, divide science into three parts. The first comprises the most theoretical principles, and those more abstract notions whose application is either unknown or very remote. The second is composed of those general truths which still belong to pure theory, but lead nevertheless by a straight and short road to practical results. Methods of application and means of execution make up the third. Each of these different portions of science may be separately cultivated, although reason and experience show that none of them can prosper long, if it be absolutely cut off from the other two.

ALEXIS DE TOCQUEVILLE,

Democracy in America

INHOUD

VOORWOORD

INLEIDING EN PROBLEEMSTELLING

I.	ENKELE ASPECTEN UIT DE COMPONENTEN- EN FACTORANALYSE	8
I.1	Inleiding	8
I.2	De principale componentenanalyse	14
I.3	De principale componenten en hun informatie	26
I.4	De componentenanalyse versus de factoranalyse	33
II.	OPTIMALE COORDINAATFUNCTIES	43
II.1	Inleiding	43
II.2	De optimale coördinaatfuncties (volgens Karhunen-Loève)	45
III.	CANONIEKE ONTWIKKELINGEN VAN STOCHASTISCHE PROCESSEN	72
III.1	Inleiding	72
III.2	De canonieke ontwikkeling van een stochastisch proces over een rij van discrete punten	73
III.3	De canonieke ontwikkeling van een stochastisch proces over een bepaald interval	84
III.4	De Karhunen-Loève ontwikkeling als een bijzonder geval van de canonieke ontwikkeling	89
	APPENDIX	94
	SUMMARY	101
	LITERATUUR	104

VOORWOORD

In 1963 publiceerde T.W. Anderson een artikel in Psychometrika onder de titel: "The use of factor analysis in the statistical analysis of multiple time series", waarin hij de suggestie deed de factoranalyse aan te wenden bij het analyseren van tijdreeksen. Anderson stelde daarbij echter nadrukkelijk dat het gebruik van de methode van de factoranalyse slechts als een eerste fase moet worden beschouwd.

De factoranalyse geeft immers alleen maar aan hoe de variabelen van een proces kunnen worden teruggebracht tot enkele, van méér fundamentele betekenis zijnde variabelen, of wel factoren.

De ladingen op deze factoren brengen de onderlinge samenhangen, welke tussen de oorspronkelijke variabelen bestaan, tot uitdrukking.

De tweede fase in Anderson's visie bestaat uit een onderzoek naar de ontwikkeling in de tijd van deze factoren, en in een onderzoek naar de relaties welke tussen deze factoren bestaan.

In bovengenoemd artikel suggereert Anderson de invoering van de tijd als een aparte factor. Deze factor dient dan te worden geëlimineerd uit de variabelen, welke deel uitmaken van het proces in kwestie. Dit betekent in feite een lineaire regressie van elke variabele op de tijd. Op de verkregen residuen dient dan daarna een factoranalyse uitgevoerd te worden.

De in dit proefschrift beschreven studie heeft de methode van de componentenanalyse als uitgangspunt.

De in theoretisch opzicht bestaande verschillen tussen het model van de factoranalyse en het model van de componentenanalyse doen geforceerd en kunstmatig aan, temeer daar de componentenanalyse als eerste aanzet voor de factoranalyse wordt gehanteerd en daarin overgaat.

Dat tussen factoren en componenten geen wezenlijk onderscheid behoeft te worden gemaakt kan ook worden aangetoond met behulp van het uit de informatie theorie bekende begrip entropie. Dit is een kwantitatieve maat voor de "onzekerheid in een systeem", waarbij

verandering in die entropie een aanwijzing is voor de aanwezige "hoeveelheid informatie". Bovenstaande overweging heeft geleid tot een voorkeur voor de componentenanalyse als methode van onderzoek, boven die van de factoranalyse.

Het tijdafhankelijke karakter van de met behulp van de methode van de componentenanalyse verkregen componenten, welke, evenals de factoren, de samenhangen tussen de oorspronkelijke variabelen tot uitdrukking brengen, wordt onderzocht in relatie tot die van de haar samenstellende variabelen.

Daartoe wordt elke variabele aan de hand van zijn tijdreeks van waarnemingen geanalyseerd. Dit gebeurt met behulp van een methode welke geheel analoog is aan die van de componentenanalyse. Zij komt neer op een ontwikkeling van de betreffende variabele in een aantal functies, welke alleen afhankelijk zijn van de tijd, en welke als componenten zijn op te vatten.

Deze methode kan als het ware als een één-dimensionale componentenanalyse worden beschouwd.

Van mijn leermeesters ben ik in de eerste plaats dank verschuldigd aan mijn promotor, Prof.Dr.J.J.J.Dalmulder.

Zijn voortdurende blijken van vertrouwen in mijn pogingen nieuwe wegen te vinden, hebben mij in mijn onderzoek, waarvan de resultaten in deze verhandeling zijn neergelegd, zeer gestimuleerd. Mijn erkentelijkheid voor Dr.J.H.F.Schilderink is niet alleen gebaseerd op het feit dat hij mij heeft geënthousiasmeerd voor de factor- en componentenanalyse. Hij heeft mij niet alleen met de mogelijkheden, maar evenzeer met de beperkingen, welke aan deze in het econometrisch onderzoek nog weinig toegepaste methode zijn verbonden, kennis laten nemen.

Voor de toewijding en precisie waarmee Anita Kuling en Marijke van den Bemt het gehele manuscript hebben getypt, kan ik niet genoeg waardering opbrengen.

Eveneens zou ik mijn waardering willen uitspreken voor de medewerking welke ik van de drukkerij en binderij van de Katholieke Hogeschool Tilburg, heb mogen ondervinden. In de persoon van de heren W.Ceron en E.Wayers zou ik al diegenen, die aan deze fase van de tot stand koming van mijn proefschrift hebben meegewerkt, van harte willen bedanken.

R.S.

INLEIDING EN PROBLEEMSTELLING

In het onderzoek naar de onderlinge samenhang tussen de grootheden van de econometrische modellen heeft de methode van de factoranalyse nog weinig ingang gevonden.

Hierbij spelen onbekendheid met de methode, alsmede een voorbijgaan aan, dan wel een in niet voldoende mate verwerken van een belangrijk aspect, namelijk de tijd, een rol.

In de factoranalyse wordt de configuratie van de factoren volkomen bepaald door de correlatiematrix R .

In het econometrisch onderzoek, wordt in overwegende mate gebruik gemaakt van tijdreeksen. Varianties en covarianties van de in het onderzoek op te nemen variabelen, berekent men dan als gemiddelden over de tijd.

De daarna bepaalde correlatiecoëfficiënten vormen de correlatiematrix R .

Past men op deze correlatiematrix een factoranalyse toe dan geven de factoren een beschrijving van de onderlinge samenhangen tussen de variabelen welke in het econometrisch model zijn opgenomen.

Deze samenhangen gelden echter alleen voor de periode welke gevormd wordt door de N tijdseenheden waarover de waarnemingen zich uitstrekken. Bij het bepalen van de varianties en covarianties als gemiddelden over de tijd, gaat men uit van de veronderstelling dat de processen welke door de in het onderzoek betrokken variabelen worden gegenereerd, stationair zijn. Deze veronderstelling is echter geenszins reëel en doet derhalve wel enige afbreuk aan de resultaten van de analyse. Zij verschaffen echter wel voldoende informatie om een eerste indruk te verkrijgen, welke als basis kan dienen voor verder onderzoek. De ideale situatie zou bestaan uit een berekening van R op de tijdstippen t , $t+1$, ..., $t+k$.

Een vergelijking van overeenkomstige factoren en ladingen, over deze verschillende situaties maakt dan een onderzoek naar de beweging van factoren en ladingen, in de tijd, mogelijk. Dit zou het dynamisch karakter van de totaalbeschrijving zeer ten goede komen. Het bezwaar dat toepassing van de methode van de factoranalyse een enigszins statische beschrijving van de structuur van het onderzoekingsgebied oplevert, is reeds door vele van haar gebruikers

onderkend.

Zo uitte Garrett[†] de veronderstelling dat het dominerende karakter van de factor -algemene intellectuele begaafdheid- met de leeftijd zou moeten afnemen, terwijl die van enkele specifieke begaafdheden juist zou moeten toenemen.

Uitgedrukt in de termen van de factoranalyse betekent dit, dat de ladingen op deze algemene factor, in de tijd gezien, kleiner zouden moeten worden, in tegenstelling tot die op de factoren der specifieke begaafdheden.

Studies van Corballis en Traub^{††} wijzen eveneens op de noodzaak het factoranalyse model zodanig aan te passen, dat wijzigingen in factoren en/of ladingen als gevolg van hun ontwikkeling in de tijd, kunnen worden gesignaleerd.

Zo stellen zij, in een artikel, gepubliceerd onder de titel -Longitudinal Factor Analysis- onder andere het volgende:

nearly all previous factor analysis of test batteries which include the same test or tests administered on more than one occasion have assumed factor scores to remain constant, which means that change is necessarily described in terms of changing factor loadings.

Models which hold factor scores constant may possess some drawbacks. Intuitively, at least, it seems unreasonable to

† H.E.A.Garret A development theory of intelligence
American Psychologist. 1946, Vol 1,
Blz. 372 - 378

†† M.C.Corballis and R.E.Traub Longitudinal Factor Analysis
Psychometrika. 1970, Vol 35, No 1,
Blz. 79 - 80.

hold factor scores constant in situations in which test scores change, for factors can be regarded, potentially at least, as tests.

The question of whether factor scores should or should not be permitted to change raises a broader issue concerning how factors should be defined in the context of change.

On the one hand, we can explicitly seek factors which represent immutable qualities of people, defining factor scores to be constant over time. This is the conventional approach. The difficulty is that there may be no such immutable qualities in any meaningful sense.

However there are some difficulties.

One is that any model which allows factor scores on the same factor to change between occasions is open to the alternative interpretation that different, though possible correlated factors are being measured on each occasion.

On the other hand, we might seek dimensions on which people may and usually do change over time.

In een C.E.S. publicatie gaven Meulepas en van Rompuy[†] als één van hun conclusies, dat:

het nuttig zou zijn de gemeenschappelijke factoren te bestuderen op verschillende tijdstippen, en de wijzigingen in de representativiteit der basisvariabelen te analyseren. Vergelijking van de resultaten van factoranalytische benaderingen van het welvaartsprofiel op verschillende tijdstippen zou aanleiding kunnen geven tot de bevestiging der hypothesen omtrent de kengetallen en van de belangrijkheid der welvaartsvariabelen in de synthese.

† E.Meulepas en P.van Rompuy Factoranalytisch onderzoek van de C.E.S. welvaartsindicatoren. Tijdschrift voor Economie 1963, No 2.

In een door Schilderlinck en van Straelen[†] met behulp van de methode van de factoranalyse verrichte studie naar de invloed van de Europese integratie op de welvaart en de economische groei in de Benelux, werd de tijd als een aparte, zelfstandige factor geïntroduceerd.

Dit gebeurde om de ontwikkeling van de Benelux economie als gevolg van de Europese integratie te kunnen onderscheiden van de trendmatige ontwikkeling die deze economie, ook zonder Europese integratie, zou ondergaan.

Hiertoe werden de tijdreeksen eerst gezuiverd van de tijd door middel van een lineaire trend.

Daar een dergelijke eliminatie van de tijd in tijdreeksen van economische grootheden zeer onvolledig is, werd vervolgens in de factoranalyse een aparte factor tijd opgenomen, waarop de variabelen werden geroteerd.

De bindingspercentages van deze factor geven dan het gedeelte van de varianties van de betreffende variabelen aan, dat door de factor tijd wordt gebonden.

Bewerkstelligt deze aanpak stellig een verbetering, zij biedt echter niet de mogelijkheid om de invloed van de Europese integratie van jaar tot jaar te kunnen vervolgen.

In de hiervoor genoemde studies spreekt men steeds van factoranalyse. Gaat men na hoe de factoren zijn bepaald, dan zal men constateren dat vrijwel steeds is uitgegaan van de methode van de principale componenten.

Uit de praktijk blijkt dat tussen factoranalyse en componentenanalyse slechts weinig verschil wordt gemaakt.

† J.H.F.Schilderlinck en

R.A. van Straelen

Proeve tot kwantitatieve analyse van de invloed der Europese integratie op de Benelux economie.

1965.

Zo stellen Lohnes en Cooley[†]:

the "construct-seeking" task of factoranalysis is most frequently accomplished today by first conducting a principal-components analysis, and by then using the resulting principal factors as a set of reference axes for determining the simplest structure, or most easily interpretable set of factors, for the domain in question.

De methode van onderzoek

Aan de door ons ontwikkelde gedachtegang voor het onderzoek naar de structuur en de samenhang van een economisch proces ligt de methode van de componentenanalyse ten grondslag.

Aangezien een toepassing hiervan slechts dan zin heeft, indien tevens het tijdafhankelijke karakter van elk der in het onderzoek betrokken variabelen in de beschouwing wordt opgenomen, vormt dit het kernpunt van onze studie.

Ons uitgangspunt wordt gevormd door de componenten waaraan getracht wordt, al dan niet na rotatie, een zinvolle interpretatie te geven. Elk zo'n component bestaat uit een lineaire combinatie van de variabelen \underline{z}_1 , \underline{z}_2 , ..., \underline{z}_p . Duidt men de i -de component aan met \underline{f}_i , dan is:

$$\underline{f}_i = b_{i1} \underline{z}_1 + b_{i2} \underline{z}_2 + \dots + b_{ij} \underline{z}_j + \dots + b_{ip} \underline{z}_p \quad (1)$$

$$(i=1,2,\dots,p)$$

[†] W.W.Cooley en P.R.Lohnes Multivariate Data Analysis
1971. Blz. 131

Hierin geeft de coëfficiënt b_{ij} van \underline{z}_j dus de mate van belangrjkheid aan van de variabele \underline{z}_j in de component \underline{f}_i .

Bovenstaande uitdrukking, welke de op de een of andere wijze tot stand gekomen onderlinge samenhang tussen de variabelen tot uitdrukking brengt, kan beschouwd worden als een model ter beschrijving van de in wezen niet meetbare component \underline{f}_i . Dit model maakt tevens deel uit van de totaalbeschrijving van de te onderzoeken structurele samenhang tussen de variabelen.

Worden nu alle variabelen \underline{z}_j uit (1) uitgedrukt als een functie van de tijd, dan is hiermee de mogelijkheid geopend om het tijdafhankelijke karakter van de betreffende component te bestuderen. Het stochastische proces $\underline{z}_j(t)$, bepaald door de stochastische variabele \underline{z}_j en de op de tijdstippen $t=1,2,\dots,N$ uitgevoerde waarnemingen, wordt daartoe weergegeven door een canonieke ontwikkeling volgens:

$$\begin{aligned}\underline{z}_j(t) &= \sum_{k=1}^n \underline{x}_{jk} \Psi_k(t) \\ &= \underline{x}_{j1} \Psi_1(t) + \underline{x}_{j2} \Psi_2(t) + \dots + \underline{x}_{jk} \Psi_k(t) + \dots + \underline{x}_{jn} \Psi_n(t) \\ &\quad (2) \\ &\quad (j=1,2,\dots,p)\end{aligned}$$

Hier zijn de alléén van de tijd afhankelijk gestelde functies $\Psi_k(t)$ niet stochastisch, in tegenstelling tot de coëfficiënten \underline{x}_{jk} , die dat wel zijn.

Uit de herschrijving van $\underline{z}_j(t)$ volgens (2), in termen van de tijdafhankelijke functies $\Psi_k(t)$ blijkt de overeenkomst in de gedachtegang welke ten grondslag ligt aan de methode van de canonieke ontwikkeling van stochastische processen en aan die van de componentenanalyse.

De componentenanalyse komt neer op een herschrijven van de variabelen \underline{z}_j in termen van de intrinsieke variabelen of componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$. De uitdrukking (2), waar de variabele \underline{z}_j wordt her-

schreven in termen van als coördinaten te beschouwen functies $\Psi_1(t)$, $\Psi_2(t)$, ..., $\Psi_n(t)$, geeft in wezen hetzelfde weer.

De in hoofdstuk II weer te geven gedachtegang betreffende de afleiding van de optimale, als coördinaten te beschouwen functies $\Phi_1(t)$, $\Phi_2(t)$, ..., $\Phi_n(t)$ verloopt geheel analoog aan die voor de componenten, zoals zal worden beschreven in de tweede paragraaf van hoofdstuk I.

Zij kan worden beschouwd als een, "één-dimensionale componenten analyse", waarbij de functie $\Phi_k(t)$ het analogon is van de component f_k waarbij $k=1,2,\dots,n$.

Onder -optimaal- dient men te verstaan, dat een maximale beschrijving wordt verkregen met een minimum aan coördinaatfuncties.

Het coördinatensysteem van functies $\{\Phi_k(t)\}$ blijkt in deze zin optimaal te zijn, indien deze functies de eigenfunctie oplossingen zijn van een homogene integraalvergelijking. De kern hiervan wordt gevormd door de correlatiefunctie van de desbetreffende variabele z_j .

In hoofdstuk III zal een methode worden ontwikkeld ter oplossing van deze integraalvergelijking.

De eigenschappen van z_j , zoals bijvoorbeeld haar trend, periodiciteiten en andere, blijken dan op een zinvolle manier in de optimale oplossing $\Phi_1(t)$, $\Phi_2(t)$, ..., verwerkt te kunnen worden.

Substitutie van (2) in (1) geeft een beschrijving van f_i in termen van de tijdafhankelijke eigenschappen van de variabelen

z_1, z_2, \dots, z_p .

Het product van b_{ij} met x_{jk} is de maat voor de importantie van de functie $\Phi_k(t)$ -behorende bij de variabele z_j -, in de component f_i . Deze "maten van belangrijkheid" kunnen vooral van belang zijn bij het toetsen van bepaalde al of niet subjectieve inzichten betreffende de tijdgevoeligheid van de variabelen, welke in het econometrisch model zijn opgenomen.

HOOFDSTUK I ENKELE ASPECTEN UIT DE COMPONENTEN- EN FACTORANALYSE

I.1 Inleiding

Als introductie voor de factoranalyse, voert Harman[†] onder andere het volgende aan:

The principle concern of factor analysis is the resolution of a set of variables linearly in terms of (usually) a small number of categories or factors.

This resolution can be accomplished by the analysis of the correlations among the variables.

A satisfactory solution will yield factors which convey all the essential information of the original set of variables. Thus the chief aim is to attain scientific parsimony or economy of description.

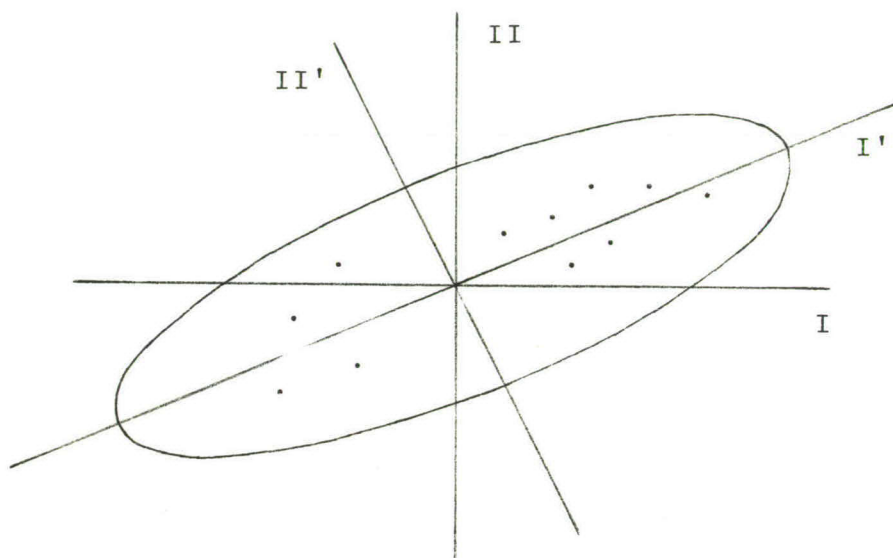
Uitgaande van het probleem waarmee omstreeks 1900, de grondlegger van de factoranalyse, Ch. Spearman, werd geconfronteerd zullen in dit hoofdstuk de door Harman voor de factoranalyse karakteristiek geachte doelstellingen worden toegelicht en uitgewerkt. Spearman's onderzoek betrof een aantal psychologische tests, voor het meten van bepaalde variabelen of eigenschappen, welke aan een N-tal personen werd voorgelegd.

Spearman constateerde correlaties tussen de scores behorende bij test I en test II; zelfs indien tussen de variabelen I en II ogenschijnlijk niets gemeenschappelijks bestond.

De verklaring die hij hiervoor gaf, bestond hieruit, dat de correlatie tussen I en II "tot stand moest zijn gekomen" via een derde variabele.

[†] H.H. Harman Modern Factor Analysis
1960, Blz. 4

Hieraan kan een geometrische interpretatie worden gegeven, door de scores (I_i , II_i) van de N proefpersonen i , als punten in een tweedimensionaal Cartesisch assenstelsel af te beelden. De puntenwolk van de scores zal dan, zoals de figuur laat zien de vorm van een ellips aannemen, waarvan het middelpunt gevormd wordt door het gemiddelde van de testcores.



Spearman's verklaring met betrekking tot de noodzakelijke aanwezigheid van een variabele, die als oorsprong kan worden beschouwd van waaruit I en II zijn ontstaan, dan wel met I en II op de een of andere wijze is verbonden, komt dan neer op het zoeken naar een coördinaatas in de tweedimensionale scoreruimte zodanig, dat de gecombineerde beschrijving van de twee eigenschappen I en II langs deze as optimaal geschiedt.

Het door Pierson[†] opgestelde en uitvoerig geargumenteerde criterium voor deze optimale beschrijving, komt er op neer, dat langs deze

† K. Pierson On lines and planes of closest fit to system of points in space.

Philosophy Magazine, 1901, 6, 559 - 572.

as de grootst mogelijke variantie in de scores (I_i , II_i) moet worden verkregen.

Dit wil niets anders zeggen, dan dat de som van de gekwadrateerde projecties van de afstanden van de punten (I_i , II_i) tot het middelpunt, op deze nieuwe as zo groot mogelijk moet zijn.

Dit wordt bereikt, indien voor deze nieuwe as de grootste hoofdas I' van de ellips wordt genomen. Langs deze hoofdas zal de door Spearman gezochte "afgeleide en niet-direct meetbare variabele of eigenschap" dienen te worden afgezet.

Deze intrinsieke eigenschap neemt aldus de grootst mogelijke variantie in de scores I_i en II_i voor haar rekening. Langs de tweede hoofdas II' van de ellips kan een tweede intrinsieke eigenschap worden afgezet die dan de resterende variantie krijgt toebedeeld.

Is deze verwaarloosbaar klein, dan kan de tweede nieuwe variabele wel worden geëlimineerd; zij levert immers toch geen wezenlijke bijdrage aan de vergroting van het inzicht in de onderlinge samenhang tussen de variabelen I en II. Worden nu de testcores van de p oorspronkelijke variabelen afgebeeld in een p-dimensionaal Cartesisch assenstelsel, dan kunnen door toepassing van het door Pearson opgestelde criterium achtereenvolgens de p verschillende hoofdassen van de door de puntenwolk van de scores gevormde p-dimensionale ellipsoïde, worden verkregen.

Deze hoofdassen- of principale componenten methode geeft dus aan tot welk intrinsieke variabelen, voortaan componenten genoemd, de oorspronkelijke variabelen, in hun totaliteit bezien, zijn terug te brengen.

Achtergrond van deze methode is de gedachte, dat door middel van deze componenten méér inzicht kan worden verkregen in de structuur van de oorspronkelijke variabelen in kwestie, dan in eerste instantie mogelijk was.

Aangezien de componenten niet direct meetbaar zijn zal dit grotere inzicht moeten worden verkregen via de relaties die er bestaan tussen oorspronkelijke variabelen en componenten; er zal, met andere woorden een model moeten worden opgesteld.

Worden de oorspronkelijke variabelen aangegeven met

\underline{z}_i ($i = 1, 2, \dots, p$) en de componenten met \underline{f}_j ($j = 1, 2, \dots, p$) dan luidt het model:

$$\underline{z}_i = a_{i1} \underline{f}_1 + a_{i2} \underline{f}_2 + \dots + a_{ip} \underline{f}_p \quad (i = 1, 2, \dots, p)$$

Elke variabele wordt dus beschreven als een lineaire combinatie van p componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$.

Deze componenten nemen bovendien achtereenvolgens een maximaal aandeel van de totale variantie van de variabele \underline{z}_i voor hun rekening.

Bij de verdere ontwikkeling van de zojuist geschetste gedachtegang en het naar aanleiding daarvan opgestelde model, staan twee wegen open:

1. waarbij alle p componenten in de beschouwing blijven betrokken.

Dit betekent een opsplitsing van de totale variantie van de \underline{z} -variabelen over alle p componenten.

2. waarbij n van de p componenten voldoende zijn.

Hierbij wordt dus de mogelijkheid opengelaten, dat sommige componenten te weinig essentiële informatie bevatten dan dat het de moeite waard zou zijn, deze voortdurend een rol in het geheel te laten blijven spelen; zij kunnen dus worden geëlimineerd.

Door deze reductie kan de analyse van het model tot hanteerbare vormen worden teruggebracht.

Een andere consequentie van deze reductie is, dat tevens een gedeelte van de totale variantie van de \underline{z} -variabelen wordt genegeerd.

In een door Lohnes en Marshall[†] verrichte studie, gepubliceerd onder de titel -Redundancy in Student Records- bleken de éénentwintig hierin betrokken variabelen teruggebracht te kunnen worden tot slechts twee componenten.

De eerste component nam 68 procent van de totale variantie voor haar rekening; de tweede component, loodrecht op de eerste, was verantwoordelijk voor 6 procent.

In beide gevallen spreekt men van componentenanalyse.

Met betrekking tot de opsplitsing van de totale variantie volgt de factoranalyse een andere weg. De totale variantie wordt hier opgesplitst in een gemeenschappelijk en in een specifiek gedeelte. De voor alle \underline{z} -variabelen gemeenschappelijke componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n$ nemen het gemeenschappelijke gedeelte van de totale variantie voor hun rekening, terwijl het specifieke gedeelte op rekening komt van een voor elke \underline{z} -variabele specifieke component. Het model voor de factoranalyse luidt:

$$\underline{z}_i = a_{i1} \underline{f}_1 + a_{i2} \underline{f}_2 + \dots + a_{in} \underline{f}_n + \underline{u}_i$$
$$(i=1,2,\dots,p)$$

Hierin is \underline{u}_i de voor de variabele \underline{z}_i specifieke component. Het aantal n van de in de factoranalyse te betrekken componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n$ dient, in tegenstelling tot de werkwijze in de componentenanalyse, van te voren te worden gespecificeerd. In de factoranalyse ondergaan de componenten in bovenstaand model tevens een naamsverandering; deze worden daar factoren genoemd.

[†] P.R.Lohnes and T.O.Marshall Redundancy in student records.
American Educational Research
Journal, 1965, 2, 19 - 23.

Analoog aan de regressieanalyse wordt in de componenten- en factoranalyse elke variabele \underline{z}_i eveneens uitgedrukt als een lineaire functie van een aantal onafhankelijke variabelen,

$$\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n, \dots, \underline{f}_p.$$

Waar echter in de regressieanalyse de onafhankelijke variabelen bepaald en waarneembaar zijn, zijn componenten en factoren hypothetische constructies, welke afgeleid dienen te worden aan de hand van de waarnemingen aan de \underline{z} -variabelen.

De in een eerder stadium gehanteerde en in nauwe relatie tot elkaar gebruikte begrippen -optimaal- en -informatie-, verdienen nog enige toelichting.

Daarvoor is in de eerste plaats quantificering van het begrip -informatie- noodzakelijk.

Fisher legde in zijn in 1942 verschenen werk -The design of experiments- al een zeker verband tussen enerzijds de variantie van een stochastische variabele, te bepalen uit een aantal waarnemingen, en anderzijds de hoeveelheid informatie die deze waarnemingen verschaffen met betrekking tot de stochastische variabele in kwestie.

Deze hoeveelheid informatie werd door Fisher gedefiniëerd als de inverse van de standaarddeviatie.

De grondlegger van de informatietheorie, Shannon[†], stelde de hoeveelheid informatie welke een waarneming aan een stochastische variabele met betrekking tot deze variabele kan verschaffen, equivalent aan de door deze waarneming veroorzaakte vermindering in onzekerheid.

In de derde paragraaf van dit hoofdstuk zal de overeenkomst worden aangetoond welke er bestaat, tussen de door Fisher en Shannon gehanteerde methoden ter quantificering van het begrip informatie. Uitgedrukt in Shannon's termen van hoeveelheden informatie komt de principale componentenanalyse dan neer op een afleiden van nieuwe variabelen, componenten, zodanig dat de eerste principale component de grootst mogelijke informatie verschaft, terwijl de volgende componenten dat in steeds afnemende mate doen.

† C.A.Shannon and W.Weaver The mathematical theory of communication.
1949

I.2. De principale componenten analyse

Alvorens aan de in I.1 gegeven introductie voor de principale componenten methode een wiskundige basis te geven, zullen eerst de volgende grootheden met hun bijbehorende dimensies worden gedefiniëerd:

\underline{z} de $(p \times 1)$ vector van gestandaardiseerde oorspronkelijke stochastische variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$,
dat wil zeggen:

$$E \quad \underline{z}_i = 0$$

(I.2.1)

$$\text{Var } \underline{z}_i = 1 \quad (i=1,2,\dots,p)$$

\underline{y} de $(p \times 1)$ vector van de principale componenten $\underline{y}_1, \underline{y}_2, \dots, \underline{y}_p$

V de $(p \times p)$ matrix waarvoor geldt dat V' de vector \underline{z} transformeert tot de vector \underline{y}

v_i de i -de, $(p \times 1)$ kolomvector van V .

v_i' transformeert \underline{z} tot \underline{y}_i .

R de $(p \times p)$ correlatie matrix van \underline{z}

Transformatie van \underline{z} zal successievelijk de p principale componenten moeten opleveren. Voor de voorlopig nog onbekende en daarom eveneens nog af te leiden transformatiematrix zal dan moeten gelden:

$$\underline{y} = V' \underline{z} \quad (I.2.2)$$

De matrix V wordt kolomsgewijs opgebouwd. We beginnen met de afleiding van de eerste kolom v_1 .

Hieraan wordt de voorwaarde opgelegd dat:

$$\underline{y}_1 = v_1' \underline{z} \quad (\text{I.2.3})$$

een maximale variantie bezit.

In verband met de éénduidigheid normeren we v_1 volgens:

$$v_1' v_1 = 1 \quad (\text{I.2.4})$$

Aangezien de variantie van \underline{y}_1 gelijk is aan:

$$\text{Var } \underline{y}_1 = E (v_1' \underline{z})^2 = E (v_1' \underline{z} \underline{z}' v_1) = v_1' R v_1$$

komt het probleem dus neer op:

het maximaliseren van $v_1' R v_1$,

onder de nevenvoorwaarde $v_1' v_1 = 1$

Toepassing van de multiplicatorenmethode van Lagrange houdt in:

het maximaliseren van $L = v_1' R v_1 - \lambda_1 (v_1' v_1 - 1)$,

waarin λ_1 de Lagrange multiplier voorstelt.

Wordt de vector van partiële afgeleiden naar de elementen van v_1 ,

$$\frac{\partial L}{\partial v_1} = 2 R v_1 - 2 \lambda_1 v_1,$$

gelijkgesteld aan nul, dan resulteert dit in de volgende relatie:

$$R v_1 = \lambda_1 v_1 \quad (I.2.5)$$

De vector v_1 is dus niets anders dan een eigenvector van de correlatiematrix R , waarbij λ_1 de bijbehorende eigenwaarde is. Wil (I.2.5) een oplossing bezitten, dan zal moeten gelden:

$$|R - \lambda I| = 0.$$

Deze polynoom in λ , van de graad p , bezit p wortels λ .

Uit:

$$\text{Var } \underline{y}_1 = v_1' R v_1 = \lambda_1 v_1' v_1 = \lambda_1 \quad (I.2.6)$$

volgt dat de bij de grootste eigenwaarde behorende eigenvector de gezochte vector v_1 is, welke tevens de vector \underline{z} transformeert tot de eerste principale component \underline{y}_1 .

De tweede kolom v_2 van V , moet nu zodanig bepaald worden dat:

$$\underline{y}_2 = v_2' \underline{z},$$

de grootste variantie bezit van alle lineaire relaties van \underline{z} , welke

niet gecorreleerd zijn met de reeds afgeleide \underline{y}_1 . Dit betekent dat:

$$\mathbf{v}_2' \mathbf{R} \mathbf{v}_2$$

gemaximaliseerd moet worden; nu onder twee nevenvoorwaarden, te weten:

$$\mathbf{v}_2' \mathbf{v}_2 = 1$$

en

$$\mathbf{v}_2' \mathbf{v}_1 = 0$$

Dit betekent weer het maximaliseren van:

$$L = \mathbf{v}_2' \mathbf{R} \mathbf{v}_2 - \lambda_2 (\mathbf{v}_2' \mathbf{v}_2 - 1) - \mu (\mathbf{v}_2' \mathbf{v}_1),$$

waarin λ_2 en μ de Lagrange multipliers voorstellen.

De vector van afgeleiden van L naar de elementen van \mathbf{v}_2 , gelijk gesteld aan nul, geeft:

$$2 \mathbf{R} \mathbf{v}_2 - 2 \lambda_2 \mathbf{v}_2 - \mu \mathbf{v}_1 = 0.$$

Vóórvermenigvuldiging van deze uitdrukking met \mathbf{v}_2' geeft

$$2 \mathbf{v}_2' \mathbf{R} \mathbf{v}_2 - 2 \lambda_2 \mathbf{v}_2' \mathbf{v}_2 - \mu \mathbf{v}_2' \mathbf{v}_1 = 0,$$

waaruit volgt:

$$v_2' R v_2 = \lambda_2.$$

Hieraan is voldaan, indien:

$$R v_2 = \lambda_2 v_2$$

Evenals naar aanleiding van (I.2.5) en (I.2.6) werd opgemerkt, geldt ook hier dat v_2 een eigenvector is van R , en dat het maximum van de variantie van \underline{y}_2 gelijk zal moeten zijn aan de op één na grootste eigenwaarde van R , en wel λ_2 .

Dit proces kan worden voortgezet tot alle p eigenvectoren en hun bijbehorende eigenwaarden zijn bepaald.

Hierna kan de transformatie matrix V worden "geformeerd"; zij bestaat uit de p naast elkaar geplaatste eigenvectoren van de correlatie matrix R van de variabelen \underline{z} .

Deze eigenvectoren corresponderen met de in afdalende grootte gerangschikte eigenwaarden van R .

De vergelijkingen ter bepaling van eigenwaarden en eigenvectoren,

$$R v_i = \lambda_i v_i \quad (i = 1, 2, \dots, p) \quad (I.2.7)$$

kunnen worden samengevat tot:

$$R V = V \Lambda \quad (I.2.8)$$

Hierin is:

$$V = (v_1, v_2, \dots, v_p), \quad (I.2.9)$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \lambda_p \end{pmatrix},$$

waarbij voor de matrix V geldt dat:

$$V' V = I.$$

Voor vermenigvuldiging van (I.2.8) met V' geeft:

$$V' R V = V' V \Lambda = \Lambda \quad (\text{I.2.10})$$

Daar:

$$|V' V| = |V'| |V| = 1$$

geldt, rekening houden met (I.2.10):

$$\begin{aligned} |R - \lambda I| &= |V'| |R - \lambda I| |V| = |V' R V - \lambda V' V| \\ &= |\Lambda - \lambda I| = \prod_{i=1}^p (\lambda_i - \lambda). \end{aligned}$$

Hieruit valt onmiddellijk af te leiden dat de wortels van $|R - \lambda I|$ gelijk zijn aan de diagonaalelementen van Λ , hetgeen

een bevestiging inhoudt van wat in een eerder stadium is afgeleid. Wordt (I.2.8) navermenigvuldigd met V^{-1} , en houdt men er rekening mede dat de eigenvectoren orthonormaal zijn, dat wil zeggen:

$$V^{-1} = V'$$

dan resulteert dit in:

$$\begin{aligned} R V V^{-1} &= R = V \Lambda V' = (V \Lambda^{\frac{1}{2}}) (V \Lambda^{\frac{1}{2}})' \\ &= \sum_{i=1}^p \lambda_i v_i v_i' \end{aligned} \tag{I.2.11}$$

R kan men zich dus opgebouwd denken als som van p afzonderlijke correlatiematrices,

$$R_j = \lambda_j v_j v_j', \quad (j = 1, 2, \dots, p) \tag{I.2.12}$$

behorende tot de principale componenten y_1, y_2, \dots, y_p .

De uitdrukking (I.2.11) betekent niets anders, dan dat door het volledige stelsel van principale componenten, de correlatiematrix R volledig wordt verklaard.

Zo kan R_1 beschouwd worden als de correlatiematrix van de oorspronkelijke waarnemingspunten (waarnemingen aan de stochastische variabelen z_1, z_2, \dots, z_p) na projectie op de eerste hoofdas. R_1 kan dus gelden als "eerste benadering" voor R, waarbij toevoeging van achtereenvolgens R_2, R_3, \dots enzovoort, voor "betere benaderingen" kan zorgen.

Worden n principale componenten voldoende geacht, dan zal R dus moeten worden opgesplitst volgens

$$R = \hat{R} + \tilde{R},$$

waarin:

$$\hat{R} = R_1 + R_2 + \dots + R_n = \sum_{i=1}^n \lambda_i v_i v_i',$$

en

$$\tilde{R} = R_{n+1} + R_{n+2} + \dots + R_p = \sum_{i=n+1}^p \lambda_i v_i v_i'$$

Aan \hat{R} kan dezelfde interpretatie worden gegeven als die voor R_1 .

Deze eerste n principale componenten tezamen, nemen van de totale variantie van \underline{z} , de grootst mogelijke hoeveelheid voor hun rekening; dus méér dan welk ander stelsel van n (genormeerde) lineaire combinaties van de variabelen \underline{z} . Voor de bepaling van de grootte van n bestaat géén objectief criterium. De enige maatstaf is, dat de geselecteerde n componenten het model op adequate wijze dienen te beschrijven.

Zo stelde D.F. Morrison[†] onder andere dat:

In practice one usually knows from earlier studies, the subject-matter nature of the data, or even the pattern of the correlations in R that a certain minimum number of components with large and distinct variances should be extracted. Beyond that number, components might be computed until some arbitrarily large proportion (perhaps 75 percent or more) of the variances has been explained. It has been my experience that if that proportion cannot be explained by the first four or five components, it is usually fruitless to persist in extracting vectors, for even if the later characteristic roots are sufficiently distinct to allow easy computation of the components, the interpretation of the components may be difficult if not possible.

[†] D.F. Morrison Multivariate Statistical Methods
1967, Blz. 228

Worden de principale componenten \underline{y}_i , ($i = 1, 2, \dots, p$), genormeerd zodanig dat hun varianties gelijk worden aan één, dan noemt men deze componenten ook wel factoren; notatie \underline{f}_i . Dit betekent, rekening houdend met (I.2.6), dat:

$$\underline{f}_i = \frac{\underline{y}_i}{\sqrt{\lambda_i}} \quad (i = 1, 2, \dots, p) \quad (\text{I.2.13})$$

Wij houden ons echter aan de benaming -componenten-.

De uitdrukkingen (I.2.13) kunnen, rekening houdend met (I.2.2) en (I.2.9), worden samengevat tot:

$$\underline{f} = \Lambda^{-\frac{1}{2}} \underline{y} = \Lambda^{-\frac{1}{2}} \underline{V}' \underline{z} = \underline{B}' \underline{z}$$

waarin:

$$\underline{B} = \underline{V} \Lambda^{-\frac{1}{2}} \quad (\text{I.2.14})$$

Het is interessant na te gaan in hoeverre de oorspronkelijke variabelen \underline{z} gecorreleerd zijn met de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$. De matrix van de correlatiecoëfficiënten noemt men de structuurmatrix; notatie \underline{S} .

De grootte van de correlatiecoëfficiënten in bijvoorbeeld de k -de kolom kunnen hulpmiddel zijn voor de interpretatie en eventueel ook voor de naamgeving van deze component. Deze correlatiecoëfficiënten geven immers weer in welke mate de variabelen

$\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ "vertegenwoordigd" zijn in de k -de component; dat wil zeggen, hoe de k -de component is "samengesteld" uit de variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$.

Anderzijds geven de correlatiecoëfficiënten in de j -de rij van \underline{S} een indruk van de mate waarin de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$ een rol spelen in de variabele \underline{z}_j .

Voor S geldt:

$$\begin{aligned} S &= E[\underline{z} \cdot \underline{f}'] = E[\underline{z} \cdot (\Lambda^{-\frac{1}{2}} V' \underline{z})'] = E[\underline{z} \cdot \underline{z}' V \Lambda^{-\frac{1}{2}}] \\ &= R V \Lambda^{-\frac{1}{2}} \end{aligned} \quad (I.2.15)$$

Gecombineerd met (I.2.8) resulteert dit in:

$$S = V \Lambda^{\frac{1}{2}}. \quad (I.2.16)$$

Evenzeer interessant is kennis met betrekking tot de coëfficiënten van de multiple regressie van de variabele \underline{z}_i op de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$; dat zijn dus de coëfficiënten a_{ij} uit de relatie,

$$\underline{z}_i = a_{i1} \underline{f}_1 + a_{i2} \underline{f}_2 + \dots + a_{ip} \underline{f}_p. \quad (I.2.17)$$

$$(i = 1, 2, \dots, p)$$

De relaties (I.2.17) kunnen worden samengevat tot:

$$\underline{z} = A \underline{f} \quad (I.2.18)$$

Navermenigvuldiging hiervan met \underline{f}' , gevolgd door de bepaling van de verwachtingswaarde van linker- en rechterlid geeft:

$$E[\underline{z} \cdot \underline{f}'] = E[A \underline{f} \cdot \underline{f}'] = A. \quad (I.2.19)$$

Immers, de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$ zijn ongecorreleerd, terwijl hun variantie door normering volgens (I.2.13) gelijk is gemaakt aan één. Uit (I.2.15) en (I.2.19) volgt, dat de matrices A en S identiek zijn.

Tevens kan men uit

$$A = S = V \Lambda^{\frac{1}{2}}, \quad (\text{I.2.20})$$

afleiden dat:

$$A' A = \Lambda^{\frac{1}{2}} V' V \Lambda^{\frac{1}{2}} = \Lambda, \quad (\text{I.2.21})$$

en dit betekent weer dat:

$$\sum_{j=1}^p a_{jk}^2 = \sum_{j=1}^p s_{jk}^2 = \lambda_k. \quad (\text{I.2.22})$$

De som van de kwadraten van de regressie coëfficiënten van de variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ op de k-de component \underline{f}_k is dus gelijk aan de k-de eigenwaarde λ_k van R.

Uit (I.2.20) volgt, rekening houdend met (I.2.11) dat:

$$A A' = V \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} V' = V \Lambda V' = R \quad (\text{I.2.23})$$

Dit betekent, gebruikmakend van (I.2.21) en (I.2.9) dat:

$$\begin{aligned} \text{Spoor } [R] &= \text{Spoor } [A A'] = \text{Spoor } [A' A] \\ &= \text{Spoor } [\Lambda] = \lambda_1 + \lambda_2 + \dots + \lambda_p \end{aligned} \quad (\text{I.2.24})$$

De som van de p eigenwaarden $\lambda_1, \lambda_2, \dots, \lambda_p$ is dus gelijk aan het spoor van R , dat wil zeggen, gelijk aan p .

Uit deze eigenschap, en uit (I.2.22) volgt dan dat:

$$\frac{\lambda_k}{p} \cdot 100$$

het percentage is van de totale aanwezige variantie van \underline{z} , dat voor rekening van de k -de component wordt genomen.

Analoog aan (I.2.22) kan de som van de kwadraten van de regressiecoëfficiënten van variabele \underline{z}_j op de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$ worden bepaald,

$$\sum_{k=1}^p a_{jk}^2 = \sum_{k=1}^p s_{jk}^2 = \text{var } \underline{z}_j = 1. \quad (\text{I.2.25})$$

Voor de eerste n principale componenten geldt dan:

$$\sum_{k=1}^n a_{jk}^2 = \sum_{k=1}^n s_{jk}^2 < 1 \quad (\text{I.2.26})$$

Deze som geeft een aanduiding voor de mate waarin de variantie in \underline{z}_j wordt verklaard door die in de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n$.

I.3 De principale componenten en hun informatie

In I.1. is een toelichting gegeven op één van de doelstellingen van de factoranalyse; namelijk, het verkrijgen van een optimale beschrijving van een in eerste instantie door de variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ bepaald systeem.

De uitdrukking -optimale beschrijving- moet in dit licht worden opgevat als het verschaffen van maximale informatie met een minimum aan (afgeleide) variabelen; dat wil zeggen, componenten. Om dit criterium te kunnen gebruiken moet het begrip informatie worden gequantificeerd. Hiertoe stelt Shannon, dat de door een waarneming aan een stochastische variabele verkregen informatie niets anders inhoudt dan een vermindering aan onzekerheid met betrekking tot deze variabele; dat wil zeggen, dat de door een waarneming opgeleverde hoeveelheid informatie wordt verkregen door de hoeveelheid die bestond voordat de waarneming werd verricht, te verminderen met de hoeveelheid onzekerheid die daarna nog bestaat. Onzekerheid met betrekking tot een gebeurtenis (een gebeurtenis is de uitkomst van een waarneming aan een stochastische variabele) is echter op haar beurt weer afhankelijk van de kans van optreden van deze gebeurtenis. Deze onzekerheid kan dus als een functie van die kans worden gedefiniëerd.

In het algemeen wordt daarvoor de volgende functie gehanteerd,

$$- {}^2 \log p,$$

waarin p de kans van optreden van de betreffende gebeurtenis voorstelt.

Komen de gebeurtenissen z_{ij} van de stochastische variabele \underline{z}_i voor met de kansen p_j zodat:

$$P [\underline{z}_i = z_{ij}] = p_j, \quad (j=1,2,\dots,N)$$

Dan bedraagt de te verwachten hoeveelheid informatie van

$$z_{i1}, z_{i2}, \dots, z_{iN} :$$

$$I = E [- \sum_j^N p_j \cdot \log p_j]$$

Voor een continu verdeelde stochastische grootheid z met dichtheidsfunctie $f(z)$ bedraagt de te verwachten hoeveelheid informatie:

$$I = - \int_{-\infty}^{+\infty} f(z) \log [f(z)] dz.$$

Is z_i normaal verdeeld met verwachtingswaarde nul, en variantie σ^2 , dan gaat deze vorm over in:

$$\begin{aligned} I &= - \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z_i^2}{2\sigma^2}} \cdot \log \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z_i^2}{2\sigma^2}} \right] dz_i \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z_i^2}{2\sigma^2}} \left[\log \sigma\sqrt{2\pi} + \frac{1}{2} \frac{z_i^2}{\sigma^2} \right] dz_i \\ &= \log \sigma\sqrt{2\pi} + \frac{1}{2} \cdot \frac{1}{\sigma^2} \cdot \sigma^2 \\ &= \frac{1}{2} [\log \sigma + \log \sqrt{2\pi} + \frac{1}{2}] \\ &= k [\log \sigma + \frac{1}{2} \log 2\pi e]. \end{aligned}$$

(I.3.1)

Hierin is:

$$k = \frac{1}{\ln 2} .$$

Is \underline{z}_i onderhevig aan storingen, $\underline{\epsilon}$, die eveneens normaal verdeeld verondersteld worden, met verwachtingswaarde nul en variantie σ_{ϵ}^2 , dan bedraagt de te verwachten hoeveelheid informatie van deze storingstermen:

$$k [\ln \sigma_{\epsilon} + \frac{1}{2} \ln 2\pi e] .$$

Aangezien de storingstermen de informatie met betrekking tot de oorspronkelijke stochastische variabele \underline{z}_i , allerminst vergroten, doch verkleinen, dat wil zeggen een geringere reductie in onzekerheid bewerkstelligen, bedraagt de uiteindelijke informatie, door de waarnemingen zèlf geleverd:

$$k [\ln \sigma - \ln \sigma_{\epsilon}] = k \ln \frac{\sigma}{\sigma_{\epsilon}} .$$

Shannon's quantificering van het begrip informatie resulteert dus in een uitdrukking welke evenredig is met de logaritmie van het omgekeerde van de standaarddeviatie σ_{ϵ} van de storingstermen ϵ op \underline{z}_i ; deze maat vertoont een grote overeenkomst met die van Fisher, die het omgekeerde van σ_{ϵ} hanteert.

Aangezien correlaties tussen en informatie met betrekking tot de variabelen van het systeem nauw met elkaar verbonden zijn, en in wezen hetzelfde tot uitdrukking brengen, zal hierop, met gebruikmaking van de juist afgeleide maat voor de te verwachten hoeveelheid informatie, nader worden ingegaan.

Aan de in de componenten analyse afgeleide componenten kan daarna een zinvolle interpretatie worden gegeven; deze is uitgedrukt in termen van -hoeveelheid informatie-.

Beschouw daartoe eerst het eenvoudigste geval; en wel dat waarbij het systeem in kwestie gekarakteriseerd is door een tweedimensionale normale verdeling, waarin de stochastische variabelen \underline{z}_1 en \underline{z}_2 beide een verwachtingswaarde nul hebben, en een variantie gelijk aan σ_1^2 respectievelijk σ_2^2 .

De voorwaardelijke variantie van \underline{z}_2 , bij gegeven \underline{z}_1 , bedraagt:

$$\text{Var} (\underline{z}_2 \mid z_1) = \sigma_{2.1}^2 = \sigma_2^2 (1 - r_{2.1}^2). \quad (\text{I.3.2})$$

Hierin is $r_{2.1}$ de correlatie coëfficiënt tussen \underline{z}_1 en \underline{z}_2 .

De te verwachten hoeveelheid informatie die \underline{z}_2 oplevert, indien bovendien gegeven is dat $\underline{z}_1 = z_1$, bedraagt, (I.3.1) en (I.3.2) in overweging genomen:

$$I_{2.1}^{(1)} = k [\ln \sigma_2 \sqrt{1-r_{2.1}^2} + \frac{1}{2} \ln 2\pi e].$$

Het tussen haakjes geplaatste getal bij I heeft betrekking op het aantal variabelen waardoor de informatie geleverd wordt.

De informatie van \underline{z}_1 en \underline{z}_2 tezamen wordt dan:

$$\begin{aligned} I^{(2)} &= I^{(1)} + I_{2.1}^{(1)} \\ &= k [\ln \sigma_1 + \frac{1}{2} \ln 2\pi e] + k [\ln \sigma_2 \sqrt{1-r_{2.1}^2} + \frac{1}{2} \ln 2\pi e] \\ &= k [\ln \sigma_1 \sigma_2 \sqrt{1-r_{2.1}^2} + \ln 2\pi e]. \end{aligned}$$

Wordt het stelsel van variabelen uitgebreid met een derde variabele \underline{z}_3 , dan is de voorwaardelijke variantie van \underline{z}_3 bij gegeven

\underline{z}_1 en \underline{z}_2 gelijk aan:

$$\text{Var} (\underline{z}_3 \mid \underline{z}_1, \underline{z}_2) = \sigma_{3.12}^2 = \sigma_3^2 (1 - r_{3.12}^2).$$

De multiële correlatiecoëfficiënt van \underline{z}_3 bij gegeven \underline{z}_1 en \underline{z}_2 wordt hier voorgesteld door $r_{3.12}$.

Uit de te verwachten hoeveelheid informatie van \underline{z}_3 , bij gegeven \underline{z}_1 en \underline{z}_2 ,

$$I_{3.12}^{(1)} = k [\ln \sigma_3 \sqrt{1-r_{3.12}^2} + \frac{1}{2} \ln 2\pi e]$$

volgt de totale hoeveelheid te verwachten informatie van \underline{z}_1 , \underline{z}_2 en \underline{z}_3 tezamen:

$$I^{(3)} = k [\ln \sigma_1 \sigma_2 \sigma_3 + \ln \sqrt{1-r_{3.12}^2} + \ln \sqrt{1-r_{2.1}^2} + \frac{3}{2} \ln 2\pi e]$$

Analoog hieraan kan de totale hoeveelheid te verwachten informatie van de variabelen \underline{z}_1 , \underline{z}_2 , ..., \underline{z}_p bepaald worden.

Deze bedraagt[†]:

$$I^{(p)} = k \left[\ln \prod_{i=1}^p \sigma_i + \frac{1}{2} \ln \prod_{i=2}^p (1 - r_{i.1.2 \dots (i-1)}^2) + \frac{p}{2} \ln 2\pi e \right]$$

$$= k \left[\ln \prod_{i=1}^p \sigma_i + \frac{1}{2} \ln |R| + \frac{p}{2} \ln 2\pi e \right],$$

[†] Zie Appendix

$$= k \left[\frac{1}{2} \ln \prod_{i=1}^p \sigma_i^2 |R| + \frac{p}{2} \ln 2\pi e \right] \quad (I.3.3)$$

waarin $|R|$ de determinant van de correlatie matrix R voorstelt.

De uitdrukking (I.3.3) vertoont dus een grote overeenkomst met de in (I.3.1) afgeleide hoeveelheid te verwachten informatie van een één dimensionale stochastische variabele.

de uitdrukking $\prod_{i=1}^p \sigma_i^2 |R|$ is het p -dimensionale analogon van de

in (I.3.1) gehanteerde variantie σ^2 ; zij wordt gegeneraliseerde variantie genoemd.

Worden de variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ genormeerd dan gaat

$\prod_{i=1}^p \sigma_i^2 |R|$ over in $|R|$, zodat (I.3.3) overgaat in:

$$I(p) = \frac{k}{2} [\ln |R| + p \ln 2\pi e] \quad (I.3.4)$$

De correlatiematrix R , volgens Ch.Harris[†]

-the numerical representation of the configuration of the variables in the variable space-

bepaalt zoals in (I.2) reeds is aangetoond, al evenzeer de configuratie van de componenten in de componenten ruimte. Dat aan de componenten nog een zinvolle betekenis kan worden toegevoegd, welke is uitgedrukt in -hoeveelheid informatie- blijkt uit de determinant van R , waarvoor, uitgaande van (I.2.11) het volgende geschreven kan worden:

† Ch.Harris Some recent developments in factor analysis. Educational and Psychological Measurement. 1964, Vol. XXIV, No 2, Blz. 194.

$$|R| = |V \Lambda V'| = |V V' \Lambda| = |V V'| |\Lambda| = |\Lambda|$$

$$= \lambda_1 \cdot \lambda_2 \cdot \dots \cdot \lambda_p$$

De uitdrukking (I.3.4) kan dan worden geschreven in de vorm:

$$I^{(p)} = \frac{k}{2} [\ln \lambda_1 + \ln 2\pi e]$$

$$+ \frac{k}{2} [\ln \lambda_2 + \ln 2\pi e]$$

$$+ \dots$$

$$+ \dots$$

$$+ \frac{k}{2} [\ln \lambda_p + \ln 2\pi e]$$

De eerste principale component, welke correspondeert met de grootste eigenwaarde λ_1 van R verschaft dus tevens de grootste hoeveelheid informatie.

De tweede principale component verschaft minder informatie dan de eerste, echter méér dan de derde, enzovoort.

De principale componenten, corresponderend met de in afnemende grootte gerangschikte eigenwaarden λ_i van R, bewerkstelligen dus eveneens een opsplitsing van de totale aanwezige hoeveelheid informatie in orthogonale componenten van afnemende grootte en belangrijkheid.

Componenten die een te geringe hoeveelheid informatie verschaffen kunnen dus om deze reden worden genegeerd.

I.4 De componentenanalyse versus de factoranalyse

De in (I.2) beschreven methode van de componentenanalyse bestaat in wezen uit een transformatie van de oorspronkelijke variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ tot een nieuw, evengroot stelsel van ongecorrleerde, afgeleide variabelen of componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$. Hierbij neemt de eerste component de grootst mogelijke hoeveelheid van de totale variantie voor haar rekening; de tweede component doet dat voor de resterende variantie enzovoort, zodanig dat alle p componenten tezamen de totale aanwezige variantie voor hun rekening nemen.

Een andere consequentie van de principale componentenmethode is de in (I.2.11) weergegeven eigenschap, namelijk de volledige opsplitsing van de correlatiematrix R in p correlatie matrices R_1, R_2, \dots, R_p , welke ieder afzonderlijk kunnen worden toegeschreven aan de componenten $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$. Hoewel een aantal, bijvoorbeeld n , van deze p componenten een groot gedeelte van de totale variantie voor hun rekening kunnen nemen, zijn echter wèl alle p componenten nodig om de correlatiematrix R volledig te kunnen verklaren. Zoals eerder is geconstateerd bestaat voor de keuze van n géén objectief criterium, maar zal men zich moeten laten leiden door zijn ervaring, met betrekking tot de materie in kwestie.

In tegenstelling tot de componentenanalyse, waar het model

$$\underline{z}_i = a_{i1} \underline{f}_1 + a_{i2} \underline{f}_2 + \dots + a_{ip} \underline{f}_p \quad (i=1,2,\dots,n,\dots,p)$$

eigenlijk het resultaat is van een bewerking, welke resulteert in een transformatie, ligt aan de factoranalyse een hypothetisch model ten grondslag.

Het model voor de factoranalyse luidt:

$$\underline{z}_i = a_{i1} \underline{f}_1 + a_{i2} \underline{f}_2 + \dots + a_{in} \underline{f}_n + \underline{u}_i \quad (I.4.1)$$

($i=1,2,\dots,n,\dots,p$)

Het model (I.4.1) impliceert (n+p) nieuwe variabelen, namelijk

$\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n, \underline{u}_1, \underline{u}_2, \dots, \underline{u}_n, \dots, \underline{u}_p$.

Hiertegenover stonden p oorspronkelijke variabelen, $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_n, \dots, \underline{z}_p$.

Dit betekent dat bepaling van de coëfficiënten a_{ik} , aan de hand van de waargenomen varianties en covarianties van de oorspronkelijke variabelen \underline{z} , allerm minst één duidig kan geschieden.

Aan de nieuwe variabelen worden de volgende voorwaarden opgelegd:

$$E \underline{f}_k = 0 \quad ; \quad \text{Cov} [\underline{f}_k \underline{f}_\ell] = 0 \quad ; \quad \text{Var} [\underline{f}_k] = 1$$

$$E \underline{u}_i = 0 \quad ; \quad \text{Cov} [\underline{u}_i \underline{u}_j] = 0 \quad \text{Var} [\underline{u}_i] = R_{u_i}$$

$$\text{Cov} [\underline{u}_i \underline{f}_k] = 0$$

$$i, j = 1, 2, \dots, n, \dots, p$$

$$k, \ell = 1, 2, \dots, n \quad (I.4.5)$$

De coëfficiënt a_{ik} van de in (I.4.3) gegeven matrix A heeft nog een bijzondere betekenis; zij is namelijk eveneens gelijk aan de covariantie tussen de variabele \underline{z}_i en de factor \underline{f}_k ($i=1, 2, \dots, n, \dots, p$; $k=1, 2, \dots, n$).

Uit (I.4.3) volgt, rekening houdend met de aan \underline{f} en \underline{u} opgelegde voorwaarden (I.4.5), dat:

$$\begin{aligned} \text{Cov} [\underline{z}, \underline{f}] &= E [(A \underline{f} + \underline{u}) \cdot \underline{f}'] - E [A \underline{f} + \underline{u}] \cdot E [\underline{f}'] \\ &= E [(A \underline{f} + \underline{u}) \cdot \underline{f}'] = A, \end{aligned} \quad (I.4.6)$$

waarmee het gestelde is aangetoond.

De coëfficiënten a_{ik} worden de factorladingen genoemd.

Uit (I.4.3) kan, rekening houdend met de voorwaarden (I.4.5), de covariantiematrix R van \underline{z} geschreven worden als:

$$R = A A' + R_u, \quad (I.4.7)$$

waarin:

$$R_u = \begin{pmatrix} Ru_1 & & & & \\ & Ru_2 & & & \\ & & \ddots & & \\ & & & Ru_n & \\ & & & & \ddots \\ 0 & & & & & Ru_p \end{pmatrix}$$

De elementen r_{ij} van R uit (I.4.7) zijn te schrijven als:

$$r_{ij} = \sum_{k=1}^n a_{ik} a_{jk} + \delta_{ij} R_{u_i} \quad (I.4.8)$$

waarin:

$$1 \quad \text{voor} \quad i = j$$

$$\delta_{ij} =$$

$$0 \quad \text{voor} \quad i \neq j$$

In het bijzonder geldt, voor i gelijk aan j , gebruikmakend van de voorwaarden (I.4.2) dat:

$$\begin{aligned} \text{Var } \underline{z}_i = r_{ii} &= \sum_{k=1}^n a_{ik}^2 + R_{u_i} \\ &= a_i^2 + R_{u_i} = 1 \end{aligned} \quad (\text{I.4.9})$$

waarin:

$$a_i^2 = \sum_{k=1}^n a_{ik}^2 \quad (\text{I.4.10})$$

De grootheid a_i^2 wordt de communaliteit van de variabele \underline{z}_i genoemd; zij geeft aan in hoeverre de variantie in de variabele \underline{z}_i door die, in de voor alle \underline{z} -variabelen gemeenschappelijke factoren

$\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n$, wordt verklaard.

De grootheid a_{ik}^2 is dus het aandeel aan de communaliteit a_i^2 van \underline{z}_i , dat geleverd wordt door de factor \underline{f}_k .

De uitdrukking (I.4.10) voor a_i^2 is het analogon van de uitdrukkingen (I.2.25) en (I.2.26) uit de paragraaf over de componentenanalyse:

$$\sum_{k=1}^p a_{ik}^2 \quad \text{respectievelijk} \quad \sum_{k=1}^n a_{ik}^2$$

Worden in de componentenanalyse alle componenten bepaald dan geldt

$$\sum_{k=1}^p a_{ik}^2 = 1$$

Vindt men achteraf dat door de eerste n componenten een voldoende groot gedeelte van de variantie van de \underline{z} -variabelen wordt verklaard, met andere woorden, beperkt men zich tot n componenten, dan geldt:

$$\sum_{k=1}^n a_{ik}^2 < 1$$

In de factoranalyse echter specificceert men n vooraf, zodat volgens (I.4.9) moet gelden:

$$\sum_{k=1}^n a_{ik}^2 = 1 - R_{u_i} < 1. \quad (I.4.11)$$

Hierbij is de variantie R_{u_i} van \underline{u}_i echter onbekend, en dient zij derhalve geschat te worden.

Reeds is geconstateerd dat bepaling van de factorladingen a_{ik} geenszins éénduidig kan geschieden.

Het niet-éénduidig karakter van de factoranalyse wordt bovendien nog versterkt door het feit dat:

1. voor de keuze van n, het aantal in model (I.4.1) op te nemen factoren, géén objectief criterium bestaat.
2. de variantie R_{u_i} van de specifieke component \underline{u}_i - waarbij $i=1,2,\dots,n,\dots,p$ - niet bekend is, en dus evenmin de communaliteit a_i^2 , die volgens (I.4.10) en (I.4.11) gelijk is aan $1 - R_{u_i}$.
3. na fixatie van n en schatting van R_{u_i} de ladingen a_{ik} nog niet éénduidig te bepalen zijn.

Wordt de matrix A namelijk navermenigvuldigd met een orthonormale matrix T dan is de covariantiematrix van \underline{z} , analoog aan (I.4.7), te schrijven als:

$$\begin{aligned} R^* &= A T (A T)' + R_u \\ &= A T T' A' + R_u \\ &= A A' + R_u \\ &= R \end{aligned}$$

Dit betekent dat, hoewel de elementen a_{ik}^* van A^T verschillen van de elementen a_{ik} van A , toch dezelfde variantie-covariantiematrix wordt verkregen, met in beide gevallen gelijke waarden voor de communaliteiten a_i^2 . ($i=1,2,\dots,n,\dots,p$) Rotatie van het stelsel van orthogonale factoren $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_n$ door middel van een orthonormale transformatiematrix T , dat wil zeggen, een wentelen van de factoren om de oorsprong waarbij de onderlinge posities van de factoren bewaard blijven, bewerkstelligt dus een andere matrix van factorladingen, waarvan de elementen a_{ik}^* ook voldoen aan de voorwaarden (I.4.8), (I.4.9) en (I.4.10).

Methodologisch gezien doet het onderscheid tussen een specifieke component \underline{u}_i en een gemeenschappelijke factor \underline{f}_k , en daarmee dus ook het onderscheid tussen factoranalyse en componentenanalyse, zeer kunstmatig aan.

De meeste praktische toepassingen, aldus Stouthard[†], en met hem onder andere Schilderinc^{††}, Cooley and Lohnes^{†††}, kunnen het best beschouwd worden als onvolledige componentenanalyses, dat wil zeggen analyses waarbij niet alle componenten worden bepaald en gebruikt. Watanabe^{††††} gaat nog een stap verder door te beweren, dat er geen enkele reden bestaat enig onderscheid te maken tussen \underline{u}_i

† Ph.C.Stouthard

Data Modellen.
dissertatie. 1965. Blz. 31

†† J.H.F.Schilderinc

Een Econometrisch Model van de
Nederlandse Economie;
een toepassing van regressie
analyse en factoranalyse.
dissertatie. 1970. Blz. 106

††† W.W.Cooley, P.R.Lohnes

Multivariate Data Analysis
1971. Blz. 131

†††† S.Watanabe

Knowing and Guessing
1970. Blz. 550

en \underline{f}_k ; ergo, dat een specifieke component in wezen een bijzonder geval is van een factor.

Ondanks de in bepaalde opzichten misschien wezenlijke verschillen tussen het model van de factoranalyse en het uit een bepaalde transformatie resulterende model van de componentenanalyse is Watanabe's bewering zeker te rechtvaardigen. Introduceer daartoe de volgende grootheid:

$$\tau_{ik} = \frac{a_{ik}^2}{\sum_{k=1}^n a_{ik}^2} \quad (\text{I.4.12})$$

Deze grootheid geeft dat gedeelte aan van de door de factoren $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_k, \dots, \underline{f}_n$ verklaarde variantie van de variabele \underline{z}_i , dat voor rekening van de factor \underline{f}_k wordt genomen.

Zo kan τ_{ik} dus als het ware beschouwd worden als de kans dat de factor \underline{f}_k "vertegenwoordigd is" in de beschrijving van de variabele \underline{z}_i .

Sommige van deze factoren \underline{f}_k zullen "gelijkkelijk over de variabelen \underline{z}_i vertegenwoordigd zijn"; met andere factoren zal dat minder het geval zijn. Een maat voor de onzekerheid met betrekking tot het al of niet "gelijkkelijk aanwezig" zijn van de factoren.

$\underline{f}_1, \underline{f}_2, \dots, \underline{f}_k, \dots, \underline{f}_n$ in de beschrijving van de variabele \underline{z}_i is de waarde van de entropiefunctie:

$$I_i = - \sum_{k=1}^n \tau_{ik}^2 \log \tau_{ik} \quad (\text{I.4.13})$$

Het maximum van I_i onder de nevenvoorwaarde:

$$\sum_{k=1}^n \tau_{ik} = 1, \quad (\text{I.4.14})$$

wordt verkregen door toepassing van de multiplicatorenmethode van Lagrange.

Dit betekent partiële differentiatie van:

$$L = - \sum_{k=1}^n \tau_{ik}^2 \log \tau_{ik} + \lambda \left(\sum_{k=1}^n \tau_{ik} - 1 \right) \quad (\text{I.4.15})$$

naar de variabelen τ_{i1} , τ_{i2} , ..., τ_{in} .

De partiële afgeleiden gelijk aan nul gesteld geeft:

$$\begin{aligned} \lambda &= \frac{1}{\ln 2} [1 + \ln \tau_{i1}] \\ \lambda &= \frac{1}{\ln 2} [1 + \ln \tau_{i2}] \\ &\vdots \\ \lambda &= \frac{1}{\ln 2} [1 + \ln \tau_{in}] \end{aligned} \quad (\text{I.4.16})$$

Deze vergelijkingen geven, rekening houdend met de nevenvoorwaarde (I.4.14) de volgende oplossingen:

$$\tau_{i1} = \tau_{i2} = \dots = \tau_{ik} = \dots = \tau_{in} = \frac{1}{n}$$

Dit betekent dat de factoren \underline{f}_1 , \underline{f}_2 , ..., \underline{f}_k , ..., \underline{f}_n even sterk zijn vertegenwoordigd in de beschrijving van de variabele \underline{z}_i .

De grootste onzekerheid in een situatie met n alternatieven treedt dus op, indien al deze alternatieven even waarschijnlijk zijn.

De maximale waarde van I_i bedraagt: $^2\log n$. Zou daarentegen bijvoorbeeld τ_{ik} gelijk zijn aan één, terwijl dan, gezien de nevenvoorwaarde (I.4.14) de andere τ 's gelijk moeten zijn aan nul, dan neemt I_i een minimale waarde aan, en wel nul.

Dat τ_{ik} gelijk is aan één, wil in wezen niets anders zeggen dan dat de factor \underline{f}_k "geheel samenvalt" met de variabele \underline{z}_i , met andere woorden, dat de factor \underline{f}_k en de variabele \underline{z}_i volkomen identiek zijn. Door rotatie van het stelsel van factoren \underline{f}_1 , \underline{f}_2 , ..., \underline{f}_k , ..., \underline{f}_n verandert de waarde van de entropie I_i , waarbij het dus kan voorkomen dat de waarde nul bereikt wordt.

In de ontstane configuratie van de factoren vallen dan de specifieke component \underline{u}_i en één der factoren samen, waardoor het gemaakte onderscheid tussen specifieke component en factor geheel kan komen te vervallen.

Samenvattend kan men dus stellen dat:

- het verschil tussen de specifieke component \underline{u}_i en de factor \underline{f}_k slechts gradueel is;
- dit verschil bepaald wordt door de mate waarin de factor \underline{f}_k vertegenwoordigd is in de beschrijving van de variabele \underline{z}_i ;
- als maat hiervoor kan dienen de waarde van de entropiefunctie I_i ; neemt I_i de waarde nul aan, dan betekent dit dat \underline{u}_i en \underline{f}_k volledig samenvallen.

We kunnen nu dus stellen, dat tussen een specifieke component en een factor geen wezenlijk onderscheid gemaakt behoeft te worden, dat wil zeggen dat er geen wezenlijk onderscheid tussen de factoranalyse en de componentenanalyse bestaat.

HOOFDSTUK II

OPTIMALE COORDINAATFUNCTIES

II.1 Inleiding

Dit hoofdstuk is er op gericht het tijdafhankelijke, stochastische proces $\underline{z}_j(t)$, optimaal te ontwikkelen in termen van, als coördinaten op te vatten functies $\Psi_1(t)$, $\Psi_2(t)$, ..., $\Psi_n(t)$, volgens:

$$\underline{z}_j(t) = \underline{x}_{j1} \Psi_1(t) + \underline{x}_{j2} \Psi_2(t) + \dots + \underline{x}_{ji} \Psi_i(t) + \dots + \underline{x}_{jn} \Psi_n(t) \\ (j=1,2,\dots,p) \quad (II.1.1)$$

De gedachtegang welke de grondslag vormt voor de afleiding van optimale, zogenaamde Karhunen-Loève functies $\Psi_1(t)$, $\Psi_2(t)$, ..., $\Psi_n(t)$ is dezelfde als die welke behoort bij de methode van de principale componenten.

Duidelijk is de analogie van (II.1.1) met het model van de componentenanalyse met n componenten:

$$z_j = a_{j1} \underline{f}_1 + a_{j2} \underline{f}_2 + \dots + a_{jn} \underline{f}_n \quad (II.1.2) \\ (j=1,2,\dots,p)$$

met dien verstande echter dat het proces $\underline{z}_j(t)$ in het model (II.1.1) ééndimensionaal is, daar het zich in de t-ruimte afspeelt, terwijl in (II.1.2) sprake is van een n-dimensionaal proces.

Het aantal coördinaatfuncties in de beschrijving (II.1.1) hoeft niet gelijk te zijn aan het aantal componenten in het model (II.1.2) van de componentenanalyse. In principe kan het aantal functies $\Psi_i(t)$ oneindig groot zijn.

Eveneens bestaat er overeenkomst tussen, enerzijds de afleiding van

de eigenwaarden λ_i en de eigenvectoren v_i uit de relaties (I.2.7), namelijk:

$$R v_i = \lambda_i v_i \quad (i=1,2,\dots,p) \quad (\text{II.1.3})$$

en anderzijds die van de Karhunen-Loève functies $\Psi_i(t)$ uit de volgende homogene integraalvergelijking:

$$\int_a^b K_{z_j}(t, t') \Psi_i(t') dt' = \lambda_i \Psi_i(t) \quad (\text{II.1.4})$$

De kern $K_{z_j}(t, t')$ van deze integraalvergelijking is de correlatiefunctie van $\underline{z}_j(t)$; λ_i de bij de eigenfunctie $\Psi_i(t)$ behorende eigenwaarde.

De vergelijking (II.1.4) is duidelijk het ééndimensionale analogon van (II.1.3).

Ter vereenvoudiging van de notatie zal in het verdere verloop van dit hoofdstuk, alswel in dat van hoofdstuk III, het stochastisch proces $\underline{z}_j(t)$ aangeduid worden door $\underline{z}(t)$, waardoor de uitdrukking (II.1.1) gelezen dient te worden als:

$$\underline{z}(t) = \underline{x}_1 \Psi_1(t) + \underline{x}_2 \Psi_2(t) + \dots + \underline{x}_i \Psi_i(t) \dots + \underline{x}_n \Psi_n(t).$$

II.2 De optimale coördinaatfuncties

(volgens Karhunen-Loève)

Een Karhunen-Loève ontwikkeling van een stochastisch proces $\underline{z}(t)$ houdt een opsplitsing in van $\underline{z}(t)$ volgens:

$$\underline{z}(t) = \sum_{i=1}^n \underline{x}_i \Psi_i(t) \quad a \leq t \leq b$$

Hierin zijn de stochastische coëfficiënten \underline{x}_i ongecorreleerd, terwijl het stelsel van orthonormale functies $\{\Psi_i(t)\}$ wordt gevormd door de eigenfunctie-oplossingen van de volgende homogene integraalvergelijking[†]:

$$\int_a^b K_z(t, t') \Psi_i(t') dt' = \lambda_i \Psi_i(t)$$

De functies $\Psi_i(t)$ zijn op te vatten als een stelsel van coördinaatassen, met behulp waarvan het proces $\underline{z}(t)$ kan worden beschreven. De grootte van de component van $\underline{z}(t)$ langs de as $\Psi_i(t)$ wordt bepaald door de grootte van \underline{x}_i ; zij is dus op te vatten als de projectie van $\underline{z}(t)$ op $\Psi_i(t)$.

Die assen $\Psi_i(t)$, waarop de projecties \underline{x}_i van $\underline{z}(t)$ klein zijn, kunnen dan, als minder belangrijk, worden genegeerd zonder dat hierdoor aan de beschrijving van $\underline{z}(t)$ te veel geweld wordt gedaan.

Evenals in de componentenanalyse zal beschrijving van $\underline{z}(t)$ door middel van een zo klein mogelijk aantal assen, gewenst zijn.

Als maat voor de belangrijkheid van de as $\Psi_i(t)$ met betrekking tot de beschrijving van $\underline{z}(t)$, kan worden gehanteerd de verwachtingswaarde van het kwadraat van de projectie van $\underline{z}(t)$ op $\Psi_i(t)$.

Indien de realisaties $z^{(\alpha)}(t)$ van $\underline{z}(t)$ plaatsvinden met de kansen $p^{(\alpha)}$ -waarbij $\alpha = 1, 2, \dots$ - dan is de verwachtingswaarde van $[x_i]^2$

[†] A.B. Baggeroer State variables and communication theory

gelijk aan:

$$\rho_i = \sum_{\alpha} p^{(\alpha)} \left[x_i^{(\alpha)} \right]^2 \quad (i=1,2,\dots,n) \quad (\text{II.2.1})$$

De verwachtingswaarden ρ_i zijn dus alle groter of gelijk aan nul. Het opleggen van een normeringsvoorwaarde, zowel aan de realisaties $z^{(\alpha)}(t)$ als aan de coördinaatfuncties $\Psi_i(t)$, resulteert in:

$$\sum_{i=1}^n \rho_i = 1,$$

waaruit volgt dat:

$$0 \leq \rho_i \leq 1. \quad (i=1,2,\dots,n)$$

Dit betekent weer dat de grootheid ρ_i kan worden opgevat als de kans dat de coördinaatfunctie $\Psi_i(t)$ "deelneemt aan de beschrijving van $\underline{z}(t)$ ".

Hoe groter deze kans is, des te belangrijker is dan de functie $\Psi_i(t)$ in de beschrijving (II.1.1).

De grootheid ρ_i is dus het analogon van de in (I.4.12) geïntroduceerde grootheid τ_{ik} .

De entropiefunctie:

$$I = - \sum_{i=1}^n \rho_i \log \rho_i$$

kan worden gehanteerd als maat voor de onzekerheid ten aanzien van het gehele stelsel van assen $\{\Psi_i(t)\}$, in het bijzonder met betrekking tot het al of niet "gelijkelijk voorkomen" van

$\Psi_1(t), \Psi_2(t), \dots, \Psi_n(t)$ in de beschrijving (II.1.1) van $\underline{z}(t)$.

Overeenkomstig een analoge afleiding als die in (I.4), bereikt I een maximum, namelijk $2 \log n$, indien:

$$\rho_1 = \rho_2 = \dots = \rho_n = \frac{1}{n}.$$

Dit betekent dat de assen $\Psi_1(t)$, $\Psi_2(t)$, ..., $\Psi_n(t)$ voor de beschrijving van $\underline{z}(t)$ even belangrijk zijn.

Is daarentegen, bijvoorbeeld ρ_j gelijk aan één, met als consequentie dat de andere ρ 's gelijk zijn aan nul, dan wil dit zeggen dat voor de beschrijving van $\underline{z}(t)$ volgens (II.1.1), slechts één enkele as, en wel $\Psi_j(t)$, nodig is.

De entropiefunctie I neemt dan een minimale waarde, en wel nul aan. Aangezien beschrijving van $\underline{z}(t)$ door middel van een minimum aantal assen de voorkeur verdient, zal dus dat stelsel van coördinaatfuncties $\{\Psi_i(t)\}$ moeten worden geselecteerd, dat de entropiefunctie zo klein mogelijk maakt. We zullen daarbij aantonen dat dit stelsel van functies, Karhunen-Loève functies zijn.

Beschouw het stelsel van n orthonormale functies $\{\Psi_i(t)\}$, gedefinieerd op het interval $a \leq t \leq b$.

Dit betekent dat:

$$\int_a^b \Psi_i(t) \Psi_j(t) dt = \delta_{ij} \quad \begin{aligned} &= 1 \text{ voor } i = j \\ &= 0 \text{ voor } i \neq j \end{aligned} \quad (II.2.2)$$

($i, j = 1, 2, \dots, n$)

Ook de realisaties $z^{(\alpha)}(t)$ worden genormeerd, en wel volgens:

$$\int_a^b [z^{(\alpha)}(t)]^2 dt = 1. \quad \alpha = 1, 2, \dots$$

De coëfficiënten $x_i^{(\alpha)}$ uit de ontwikkeling van $z^{(\alpha)}(t)$,

$$z^{(\alpha)}(t) = \sum_{i=1}^n x_i^{(\alpha)} \psi_i(t), \quad (\text{II.2.3})$$

kunnen dan, onder gebruikmaking van (II.2.2), bepaald worden volgens:

$$x_i^{(\alpha)} = \int_a^b z^{(\alpha)}(t) \psi_i(t) dt \quad (\text{II.2.4})$$

Behalve $\{\psi_i(t)\}$ beschouwen we nu een ander stelsel van orthonormale coördinaatfuncties, $\{\phi_j(t)\}$, om vervolgens $z^{(\alpha)}(t)$ in dit nieuwe stelsel te beschrijven:

$$z^{(\alpha)}(t) = \sum_{j=1}^n y_j^{(\alpha)} \phi_j(t) \quad (\text{II.2.5})$$

De coëfficiënten $y_j^{(\alpha)}$ hierin zijn dan gelijk aan:

$$y_j^{(\alpha)} = \int_a^b z^{(\alpha)}(t) \phi_j(t) dt \quad (\text{II.2.6})$$

Substitutie van (II.2.3) in (II.2.6) geeft:

$$\begin{aligned} y_j^{(\alpha)} &= \int_a^b \sum_{i=1}^n x_i^{(\alpha)} \psi_i(t) \phi_j(t) dt \\ &= \sum_{i=1}^n x_i^{(\alpha)} \int_a^b \psi_i(t) \phi_j(t) dt \end{aligned}$$

$$= \sum_{i=1}^n x_i^{(\alpha)} c_{ij} , \quad (j=1,2,\dots,n) \quad (\text{II.2.7})$$

$$\text{waarin: } c_{ij} = \int_a^b \psi_i(t) \phi_j(t) dt$$

Evenzo geeft substitutie van (II.2.5) in (II.2.4):

$$x_i^{(\alpha)} = \int_a^b \sum_{j=1}^n y_j^{(\alpha)} \phi_j(t) \psi_i(t) dt$$

$$= \sum_{j=1}^n y_j^{(\alpha)} \int_a^b \phi_j(t) \psi_i(t) dt$$

$$= \sum_{j=1}^n y_j^{(\alpha)} c_{ji} , \quad (i=1,2,\dots,n) \quad (\text{II.2.8})$$

$$\text{waarin: } c_{ji} = \int_a^b \phi_j(t) \psi_i(t) dt$$

Dit betekent dat:

$$c_{ij} = c_{ji} = \int_a^b \psi_i(t) \phi_j(t) dt \quad (\text{II.2.9})$$

Door (II.2.8) in (II.2.7) en (II.2.7) in (II.2.8) te substitueren krijgen we, rekening houdend met (II.2.9), achtereenvolgens:

$$y_j^{(\alpha)} = \sum_{i=1}^n c_{ij} \sum_{j'=1}^n y_{j'}^{(\alpha)} c_{j'i} = \sum_{j'=1}^n \sum_{i=1}^n c_{ji} c_{ij'} y_{j'}^{(\alpha)}$$

$$x_i^{(\alpha)} = \sum_{j=1}^n c_{ji} \sum_{i'=1}^n x_{i'}^{(\alpha)} c_{i'j} = \sum_{i'=1}^n \sum_{j=1}^n c_{ij} c_{ji'} x_{i'}^{(\alpha)}$$

Dit resulteert in de volgende betrekkingen:

$$\sum_{i=1}^n c_{ji} c_{ij'} = \delta_{jj'} = \begin{cases} 1 & \text{voor } j = j' \\ 0 & \text{voor } j \neq j' \end{cases} \quad (\text{II.2.10})$$

$$\sum_{j=1}^n c_{ij} c_{ji'} = \delta_{ii'} = \begin{cases} 1 & \text{voor } i = i' \\ 0 & \text{voor } i \neq i' \end{cases} \quad (\text{II.2.11})$$

Wordt (II.2.7) in matrix vorm geschreven,

$$y^{(\alpha)} = C x^{(\alpha)} \quad (\text{II.2.12})$$

waarin:

$$y^{(\alpha)} = \begin{pmatrix} y_1^{(\alpha)} \\ y_2^{(\alpha)} \\ \vdots \\ y_n^{(\alpha)} \end{pmatrix}; \quad C = \begin{pmatrix} c_{11} & c_{12} \cdots c_{1n} \\ c_{21} & c_{22} \cdots c_{2n} \\ \vdots & \vdots \\ c_{n1} & c_{n2} \cdots c_{nn} \end{pmatrix}; \quad x^{(\alpha)} = \begin{pmatrix} x_1^{(\alpha)} \\ x_2^{(\alpha)} \\ \vdots \\ x_n^{(\alpha)} \end{pmatrix}$$

dan is (II.2.8) te schrijven als:

$$\mathbf{x}^{(\alpha)} = \mathbf{C}^{-1} \mathbf{y}^{(\alpha)}. \quad (\text{II.2.13})$$

Uit (II.2.9), (II.2.10) en (II.2.11) volgt dan:

$$\mathbf{C} = \mathbf{C}^T = \mathbf{C}^{-1} \quad (\text{II.2.14})$$

De stelsels coördinaatassen $\{\Psi_i(t)\}$ en $\{\Phi_j(t)\}$ zijn dus via een orthogonale transformatie matrix \mathbf{C} met elkaar verbonden.

Hier dringt zich een vergelijking op met de in (I.4) geïntrodeerde orthonormale transformatiematrix \mathbf{T} . Deze matrix bewerkstelligt een rotatie van het stelsel van orthogonale factoren, zoals de matrix \mathbf{C} uit (II.2.14) dat doet met het stelsel van orthogonale coördinaatfuncties $\{\Psi_i(t)\}$, naar het stelsel $\{\Phi_j(t)\}$.

De communaliteit van $\underline{z}(t)$ in de $\{\Psi_i(t)\}$ -ontwikkeling volgens (II.2.3), is gelijk aan:

$$\sum_{i=1}^n [x_i^{(\alpha)}]^2.$$

In de $\{\Phi_j(t)\}$ -ontwikkeling volgens (II.2.5) is de communaliteit van $\underline{z}(t)$ gelijk aan:

$$\sum_{j=1}^n [y_j^{(\alpha)}]^2.$$

Evenals dat in de factoranalyse het geval is zal door rotatie van het stelsel $\{\Psi_i(t)\}$ naar het stelsel $\{\Phi_j(t)\}$ de communaliteit niet veranderen.

Uit (II.2.7) volgt namelijk, rekening houdend met (II.2.9), (II.2.10) en (II.2.11) dat:

$$\begin{aligned}
 \sum_{j=1}^n [y_j^{(\alpha)}]^2 &= \sum_{j=1}^n \left[\sum_{i=1}^n c_{ji} x_i^{(\alpha)} \right]^2 \\
 &= \sum_{j=1}^n \sum_{i=1}^n c_{ji}^2 [x_i^{(\alpha)}]^2 \\
 &\quad + \sum_{j=1}^n \sum_{\substack{i=1 \\ i \neq i'}}^n \sum_{i'=1}^n c_{ji} c_{ji'} x_i^{(\alpha)} x_{i'}^{(\alpha)} \\
 &= \sum_{i=1}^n [x_i^{(\alpha)}]^2
 \end{aligned}$$

De autocorrelatie functie.

De autocorrelatie functie van $\underline{z}(t)$ is gedefiniëerd als:

$$K_z(t, t') = E [\underline{z}(t) \underline{z}(t')] = \sum_{\alpha} p^{(\alpha)} z^{(\alpha)}(t) z^{(\alpha)}(t') \quad (\text{II.2.16})$$

Wordt $z^{(\alpha)}(t)$ ontwikkeld volgens:

$$z^{(\alpha)}(t) = \sum_{i=1}^n x_i^{(\alpha)} \psi_i(t) \quad (\text{II.2.17})$$

dan is (II.2.16) te herschrijven tot:

$$K_z(t, t') = \sum_{\alpha} p^{(\alpha)} \sum_{i=1}^n x_i^{(\alpha)} \psi_i(t) \sum_{i'=1}^n x_{i'}^{(\alpha)} \psi_{i'}(t')$$

$$\begin{aligned}
 &= \sum_{i=1}^n \sum_{i'=1}^n \Psi_i(t) \sum_{\alpha} p^{(\alpha)} x_i^{(\alpha)} x_{i'}^{(\alpha)} \Psi_{i'}(t') \\
 &= \sum_{i=1}^n \sum_{i'=1}^n \Psi_i(t) X_{ii'}^{(\alpha)} \Psi_{i'}(t') \quad (\text{II.2.18})
 \end{aligned}$$

waarin:

$$X_{ii'}^{(\alpha)} = \sum_{\alpha} p^{(\alpha)} x_i^{(\alpha)} x_{i'}^{(\alpha)} \quad (\text{II.2.19})$$

De ρ_i 's, gedefiniëerd in (II.2.1) als:

$$\rho_i = \sum_{\alpha} p^{(\alpha)} [x_i^{(\alpha)}]^2$$

zijn dus de diagonaalelementen van de als een matrix op te vatten $X_{ii'}^{(\alpha)}$.

Zijn de coëfficiënten $x_i^{(\alpha)}$ uit (II.2.17) karakteristiek voor de beschrijving van $z^{(\alpha)}(t)$ in het $\{\Psi_i(t)\}$ - coördinaatassenstelsel, de in matrixvorm te schrijven en voortaan als matrix op te vatten $X_{ii'}^{(\alpha)}$, is dat voor de bij $\underline{z}(t)$ behorende autocorrelatiefunctie $K_z(t, t')$.

Deze vorm is, onder gebruikmaking van (II.2.4) en (II.2.16) te herleiden tot:

$$\begin{aligned}
 X_{ii'}^{(\alpha)} &= \sum_{\alpha} p^{(\alpha)} \int_a^b z^{(\alpha)}(t) \Psi_i(t) dt \int_a^b z^{(\alpha)}(t') \Psi_{i'}(t') dt' \\
 &= \int_a^b \int_a^b \sum_{\alpha} p^{(\alpha)} z^{(\alpha)}(t) z^{(\alpha)}(t') \Psi_i(t) \Psi_{i'}(t') dt dt'
 \end{aligned}$$

$$= \int_a^b \int_a^b K_z(t, t') \Psi_i(t) \Psi_{i'}(t') dt dt' \quad (\text{II.2.20})$$

Wordt $z^{(\alpha)}(t)$ ontwikkeld met behulp van het stelsel $\{\Phi_j(t)\}$, in plaats van met het stelsel $\{\Psi_i(t)\}$, en wel volgens:

$$z^{(\alpha)}(t) = \sum_{j=1}^n y_j^{(\alpha)} \Phi_j(t) \quad (\text{II.2.21})$$

dan krijgen we, de volgende aan $X_{ii'}^{(\alpha)}$ analoge vorm:

$$Y_{jj'}^{(\alpha)} = \sum_{\alpha} p^{(\alpha)} y_j^{(\alpha)} y_{j'}^{(\alpha)}$$

Deze vorm is onder gebruikmaking van (II.2.7), (II.2.9) en (II.2.19), te herleiden tot:

$$\begin{aligned} & \sum_{\alpha} p^{(\alpha)} \sum_{i=1}^n c_{ji} x_i^{(\alpha)} \sum_{i'=1}^n c_{j'i'} x_{i'}^{(\alpha)} \\ &= \sum_{i=1}^n \sum_{i'=1}^n c_{ji} \sum_{\alpha} p^{(\alpha)} x_i^{(\alpha)} x_{i'}^{(\alpha)} c_{i'j'} \\ &= \sum_{i=1}^n \sum_{i'=1}^n c_{ji} X_{ii'}^{(\alpha)} c_{i'j'} \quad (\text{II.2.22}) \end{aligned}$$

Ook $Y_{jj'}^{(\alpha)}$ en $X_{ii'}^{(\alpha)}$ zijn dus verbonden via de orthogonale transformatiematrix C uit (II.2.14), en wel als volgt:

$$Y_{jj'}^{(\alpha)} = C X_{ii'}^{(\alpha)} C'.$$

Eigenfuncties en eigenwaarden.

Veronderstel dat er een functie $\Phi_k(t)$ en een constante λ_k bestaan, hierbij is de index k voorlopig nog vast, zodanig dat aan de volgende relatie is voldaan:

$$\int_a^b K_z(t, t') \Phi_k(t') dt' = \lambda_k \Phi_k(t) \quad (\text{II.2.23})$$

Aan de functie $\Phi_k(t)$ wordt bovendien de normeringsvoorwaarde opgelegd, met andere woorden:

$$\int_a^b [\Phi_k(t)]^2 dt = 1 \quad (\text{II.2.24})$$

$\Phi_k(t)$ noemt men een eigenfunctie oplossing van (II.2.23); λ_k de bij haar behorende eigenwaarde.

Zou de functie $\Phi_k(t)$ ontwikkeld kunnen worden in de functies $\{\Psi_i(t)\}$ volgens:

$$\Phi_k(t) = \sum_{i=1}^n d_{ki} \Psi_i(t), \quad (\text{II.2.25})$$

dan moeten de coëfficiënten d_{ki} , naar analogie van (II.2.4), gelijk zijn aan:

$$d_{ki} = \int_a^b \Phi_k(t) \Psi_i(t) dt \quad (\text{II.2.26})$$

Uit (II.2.9) volgt dan dat:

$$d_{ki} = d_{ik} = c_{ik} = c_{ki} \quad (\text{II.2.27})$$

Zo gaat (II.2.23), onder gebruikmaking van (II.2.18), (II.2.25), (II.2.26) en (II.2.27) over in:

$$\begin{aligned} \int_a^b K_z(t, t') \Phi_k(t') dt' &= \int_a^b \sum_{i=1}^n \sum_{i'=1}^n \Psi_i(t) X_{ii'}^{(\alpha)} \Psi_{i'}(t') \Phi_k(t') dt' \\ &= \sum_{i=1}^n \sum_{i'=1}^n \Psi_i(t) X_{ii'}^{(\alpha)} \int_a^b \Psi_{i'}(t') \Phi_k(t') dt' \\ &= \sum_{i=1}^n \sum_{i'=1}^n \Psi_i(t) X_{ii'}^{(\alpha)} d_{i',k} \\ &= \lambda_k \sum_{i=1}^n d_{ki} \Psi_i(t) \end{aligned}$$

Hieruit is weer af te leiden, rekening houdend met de symmetrie-eigenschap (II.2.27) van de coëfficiënten d_{ki} , dat:

$$\sum_{i'=1}^n X_{ii'}^{(\alpha)} d_{i'k} = \lambda_k d_{ik} \quad (\text{II.2.28})$$

$$(i=1,2,\dots,n)$$

Deze vergelijkingen kunnen dus beschouwd worden als een stelsel van homogene lineaire vergelijkingen ter oplossing van de onbekenden d_{ik} . (k is nog steeds vast).

Wordt de zojuist uitgevoerde procedure, welke resulteerde in (II.2.28), herhaald, nu echter voor de eigenfunctie $\Phi_k(t)$ met bijbehorende eigenwaarde λ_k , dan wordt een aan (II.2.28) analoge uitdrukking verkregen:

$$\sum_{i'=1}^n d_{k'i'} X_{i'i}^{(\alpha)} = \lambda_{k'} d_{k'i} \quad (\text{II.2.29})$$

Voorvermenigvuldiging van (II.2.28) met $d_{k'i}$, gevolgd door een sommatie over i , resulteert dan, na gebruikmaking van de symmetrie-eigenschappen van $X_{ii'}^{(\alpha)}$ en d_{ik} , in:

$$\sum_{i=1}^n \sum_{i'=1}^n d_{k'i} X_{ii'}^{(\alpha)} d_{i'k} = \lambda_k \sum_{i=1}^n d_{k'i} d_{ik}$$

Zo levert navermenigvuldiging van (II.2.29) met d_{ik} , gevolgd door een sommatie over i :

$$\sum_{i=1}^n \sum_{i'=1}^n d_{k'i'} X_{i'i}^{(\alpha)} d_{ik} = \lambda_{k'} \sum_{i=1}^n d_{k'i} d_{ik}$$

Worden bovenstaande vormen van elkaar afgetrokken dan geeft dit:

$$(\lambda_k - \lambda_{k'}) \sum_{i=1}^n d_{k'i} d_{ik} = 0,$$

waaruit kan worden geconcludeerd, dat indien λ_k ongelijk is aan $\lambda_{k'}$, dan moet gelden:

$$\sum_{i=1}^n d_{k'i} d_{ik} = 0 \quad (\text{II.2.30})$$

Voor het geval dat λ_k ongelijk is aan $\lambda_{k'}$, kunnen we bewijzen, dat de bijbehorende eigenfuncties $\Phi_k(t)$ en $\Phi_{k'}(t)$ orthogonaal zijn, met andere woorden:

$$\int_a^b \Phi_k(t) \Phi_{k'}(t) dt = 0$$

Uit (II.2.23) volgt namelijk, door vermenigvuldiging van linker- en rechterlid met $\Phi_{k'}(t)$ gevolgd door een integratie over het interval $[a, b]$ dat:

$$\int_a^b \Phi_{k'}(t) \left[\int_a^b K_z(t, t') \Phi_k(t') dt' \right] dt = \lambda_k \int_a^b \Phi_{k'}(t) \Phi_k(t) dt \quad (\text{II.2.31})$$

Door op de aan (II.2.23) analoge uitdrukking:

$$\int_a^b K_z(t, t') \Phi_{k'}(t') dt' = \lambda_{k'} \Phi_{k'}(t)$$

een soortgelijke procedure toe te passen, namelijk vermenigvuldiging van linker- en rechterlid met $\Phi_k(t)$ gevolgd door een integratie over het interval $[a, b]$, krijgen we:

$$\int_a^b \Phi_k(t) \left[\int_a^b K_z(t, t') \Phi_{k'}(t') dt' \right] dt = \lambda_{k'} \int_a^b \Phi_k(t) \Phi_{k'}(t) dt. \quad (\text{II.2.32})$$

Krachtens de definitie in (II.2.16) geldt:

$$K_z(t, t') = K_z(t', t)$$

Gebruikmakend van deze eigenschap is (II.2.31) te schrijven als:

$$\int_a^b \Phi_k(t') \left[\int_a^b K_z(t', t) \Phi_{k'}(t) dt \right] dt' = \lambda_k \int_a^b \Phi_{k'}(t) \Phi_k(t) dt \quad (\text{II.2.33})$$

Aangezien het linkerlid van (II.2.32) gelijk is aan dat van (II.2.33), krijgen we, door de uitdrukkingen (II.2.32) en (II.2.33) van elkaar af te trekken de volgende vorm:

$$(\lambda_k - \lambda_{k'}) \int_a^b \Phi_k(t) \Phi_{k'}(t) dt = 0.$$

Indien nu λ_k ongelijk is aan $\lambda_{k'}$, betekent dit dat:

$$\int_a^b \Phi_k(t) \Phi_{k'}(t) dt = 0. \quad (\text{II.2.34})$$

waarmee de orthogonaliteit van de eigenfuncties $\Phi_k(t)$ en $\Phi_{k'}(t)$ is bewezen.

De eigenschappen van $\Phi_k(t)$, weergegeven in (II.2.24) en (II.2.34) kunnen nu samengevat worden tot:

$$\int_a^b \Phi_k(t) \Phi_{k'}(t) dt = \delta_{kk'} = \begin{cases} 1 & \text{voor } k = k' \\ 0 & \text{voor } k \neq k' \end{cases} \quad (\text{II.2.35})$$

De uitdrukking (II.2.30) betekent in feite, gezien (II.2.27) en (II.2.10), dat:

$$\sum_{i=1}^n d_{ki} d_{ik'} = \delta_{kk'} \quad (k, k' = 1, 2, \dots, n) \quad (\text{II.2.36})$$

Overeenkomstig een zelfde redenering als die welke leidde tot de relatie (II.2.13), kan het stelsel $\{\Phi_k(t)\}$, bestaande uit de eigenfuncties van de integraalvergelijking (II.2.23), beschouwd worden als een ander stelsel van orthogonale coördinaatassen. Dit stelsel blijkt een aantal interessante eigenschappen te bezitten.

Voor het onderzoek hiernaar, in het bijzonder in relatie tot die, welke behoren bij een willekeurig ander, bijvoorbeeld $\{\Psi_i(t)\}$ - coördinaatassenstelsel, ontwikkelen we $z^{(\alpha)}(t)$ volgens:

$$z^{(\alpha)}(t) = \sum_{k=1}^n c_k^{(\alpha)} \Phi_k(t) \quad (\text{II.2.37})$$

De coëfficiënten $c_k^{(\alpha)}$ zijn te bepalen volgens:

$$c_k^{(\alpha)} = \int_a^b z^{(\alpha)}(t) \Phi_k(t) dt. \quad (\text{II.2.38})$$

Substitueren we hierin voor $z^{(\alpha)}(t)$ de vorm (II.2.3),

$$z^{(\alpha)}(t) = \sum_{i=1}^n x_i^{(\alpha)} \Psi_i(t),$$

dat is dus $z^{(\alpha)}(t)$ uitgedrukt in het $\{\Psi_i(t)\}$ -coördinaatassenstelsel,

dan volgt hieruit, rekening houdend met (II.2.26) en (II.2.27), de relatie tussen $c_k^{(\alpha)}$ en $x_i^{(\alpha)}$:

$$\begin{aligned}
 c_k^{(\alpha)} &= \int_a^b z^{(\alpha)}(t) \Phi_k(t) dt = \int_a^b \sum_{i=1}^n x_i^{(\alpha)} \Psi_i(t) \Phi_k(t) dt \\
 &= \sum_{i=1}^n x_i^{(\alpha)} \int_a^b \Psi_i(t) \Phi_k(t) dt \\
 &= \sum_{i=1}^n x_i^{(\alpha)} d_{ik}.
 \end{aligned} \tag{II.2.39}$$

Indien $z^{(\alpha)}(t)$ wordt ontwikkeld volgens (II.2.37), dan betekent dit voor de autocorrelatiefunctie (II.2.16) dat:

$$\begin{aligned}
 K_z(t, t') &= \sum_{\alpha} p^{(\alpha)} z^{(\alpha)}(t) z^{(\alpha)}(t') \\
 &= \sum_{k=1}^n \sum_{k'=1}^n \Phi_k(t) \sum_{\alpha} p^{(\alpha)} c_k^{(\alpha)} c_{k'}^{(\alpha)} \Phi_{k'}(t) \\
 &= \sum_{k=1}^n \sum_{k'=1}^n \Phi_k(t) z_{kk'}^{(\alpha)} \Phi_{k'}(t) \\
 &= \sum_{k=1}^n \sum_{k'=1}^n \Phi_k(t) \lambda_{k'} \delta_{kk'} \Phi_{k'}(t)
 \end{aligned} \tag{II.2.40}$$

Immers uit (II.2.39) en (II.2.19) volgt:

$$\begin{aligned}
 z_{kk'}^{(\alpha)} &= \sum_{\alpha} p^{(\alpha)} c_k^{(\alpha)} c_{k'}^{(\alpha)} = \sum_{\alpha} p^{(\alpha)} \sum_{i=1}^n x_i^{(\alpha)} d_{ik} \sum_{i'=1}^n x_{i'}^{(\alpha)} d_{i'k'} \\
 &= \sum_{i'=1}^n \sum_{i=1}^n d_{ik} \sum_{\alpha} p^{(\alpha)} x_i^{(\alpha)} x_{i'}^{(\alpha)} d_{i'k'} \\
 &= \sum_{i=1}^n d_{ik} \sum_{i'=1}^n x_{ii'}^{(\alpha)} d_{i'k'}
 \end{aligned}$$

welke analoog is aan (II.2.22).

Deze vorm gaat, rekening houdend met (II.2.28) en (II.2.36) over in:

$$\begin{aligned}
 &\sum_{i=1}^n d_{ik} \lambda_k d_{ik'} \\
 &= \lambda_k \sum_{i=1}^n d_{ki} d_{ik'} \\
 &= \lambda_k \delta_{kk'} \quad (\text{II.2.41})
 \end{aligned}$$

De uitdrukking (II.2.40), met andere woorden:

$$K_z(t, t') = \sum_{k=1}^n \sum_{k'=1}^n \phi_k(t) \lambda_k \delta_{kk'} \phi_{k'}(t),$$

kan worden herschreven tot:

$$\begin{pmatrix} \Phi_1(t) & \Phi_2(t) & \dots & \Phi_n(t) \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \begin{pmatrix} \Phi_1(t) \\ \Phi_2(t) \\ \vdots \\ \Phi_n(t) \end{pmatrix} \\ = \begin{pmatrix} \sqrt{\lambda_1} \Phi_1(t) & \sqrt{\lambda_2} \Phi_2(t) & \dots & \sqrt{\lambda_n} \Phi_n(t) \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} \Phi_1(t) \\ \sqrt{\lambda_2} \Phi_2(t) \\ \vdots \\ \sqrt{\lambda_n} \Phi_n(t) \end{pmatrix} \quad (\text{II.2.42})$$

Deze vorm lijkt bijzonder veel op (I.2.23), namelijk:

$$R = A A',$$

waarbij $\begin{pmatrix} \sqrt{\lambda_1} \Phi_1(t) & \sqrt{\lambda_2} \Phi_2(t) & \dots & \sqrt{\lambda_n} \Phi_n(t) \end{pmatrix}$ als het analogon van de coëfficiënten matrix A kan worden beschouwd.

Het analogon van de uitdrukking (I.2.21), namelijk:

$$\Lambda = A' A,$$

is uit (II.2.42) af te leiden,

$$\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} \Phi_1(t) \\ \sqrt{\lambda_2} \Phi_2(t) \\ \vdots \\ \sqrt{\lambda_n} \Phi_n(t) \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1} \Phi_1(t) & \sqrt{\lambda_2} \Phi_2(t) & \dots & \sqrt{\lambda_n} \Phi_n(t) \end{pmatrix} \quad (\text{II.2.43})$$

Hierbij dient men voor het product $\Phi_k(t) \Phi_{k'}(t)$ dus eigenlijk de integraal hierover in gedachten te houden, en daarbij rekening te houden met het feit (II.2.35) dat:

$$\int_a^b \Phi_k(t) \Phi_{k'}(t) dt = \delta_{kk'},$$

Uit (II.2.41) volgt dus, dat indien $\underline{z}(t)$ wordt beschreven in het coördinaatassenstelsel $\{\Phi_k(t)\}$, de uitdrukking $Z_{kk'}^{(\alpha)}$ kan worden geschreven in de vorm van een diagonaalmatrix, met de eigenwaarden λ_k op de hooddagonaal. Deze eigenwaarden kunnen uit (II.2.41) worden gevonden door k gelijk te stellen aan k' ,

$$\lambda_k = \sum_{\alpha} p^{(\alpha)} \left[c_k^{(\alpha)} \right]^2 \geq 0 \quad (\text{II.2.44})$$

Er zij aan herinnerd dat de functies $\Phi_k(t)$ de eigenfunctie-oplossingen en λ_k de daarbij behorende eigenwaarden zijn van de integraalvergelijking:

$$\int_a^b K_z(t, t') \Phi_k(t') dt' = \lambda_k \Phi_k(t).$$

Indien we aan de realisaties $z^{(\alpha)}(t)$ van $\underline{z}(t)$ de normeringsvoorwaarde:

$$\int_a^b \left[z^{(\alpha)}(t) \right]^2 dt = 1, \quad (\text{II.2.45})$$

opleggen, dan is te bewijzen dat:

$$\sum_{k=1}^n \lambda_k = 1. \quad (\text{II.2.46})$$

Dit stemt geheel overeen met (I.2.24), met dien verstande echter dat in (I.2.24) de som van de eigenwaarden gelijk is aan p , dat wil zeggen,

gelijk is aan het aantal variabelen z_1, z_2, \dots, z_p welke betrokken zijn in de componentenanalyse.

Het aantal variabelen in de ontwikkeling (II.1.1) daarentegen bedraagt één; dat is namelijk de variabele $z_j(t)$.

Voor het bewijs van (II.2.46) substitueren we (II.2.37) in (II.2.45).

Dit geeft:

$$\begin{aligned}
 \int_a^b [z^{(\alpha)}(t)]^2 dt &= \int_a^b \sum_{k=1}^n c_k^{(\alpha)} \phi_k(t) \sum_{k'=1}^n c_{k'}^{(\alpha)} \phi_{k'}(t) dt \\
 &= \sum_{k=1}^n \sum_{k'=1}^n c_k^{(\alpha)} c_{k'}^{(\alpha)} \int_a^b \phi_k(t) \phi_{k'}(t) dt \\
 &= \sum_{k=1}^n \sum_{k'=1}^n c_k^{(\alpha)} c_{k'}^{(\alpha)} \delta_{kk'} \\
 &= \sum_{k=1}^n [c_k^{(\alpha)}]^2 = 1 \quad (\text{II.2.47}) \\
 &\quad (\alpha=1, 2, \dots)
 \end{aligned}$$

Uit (II.2.44) volgt nu, rekening houdend met (II.2.47):

$$\begin{aligned}
 \sum_{k=1}^n \lambda_k &= \sum_{k=1}^n \sum_{\alpha} p^{(\alpha)} [c_k^{(\alpha)}]^2 \\
 &= p^{(1)} \sum_{k=1}^n [c_k^{(1)}]^2 + p^{(2)} \sum_{k=1}^n [c_k^{(2)}]^2 + \dots + \dots \\
 &= p^{(1)} + p^{(2)} + \dots + p^{(n)} \\
 &= 1,
 \end{aligned}$$

waarmee het gestelde is bewezen .

Voor de ρ_i 's, volgens (II.2.1) gelijk aan:

$$\rho_i = \sum_{\alpha} p^{(\alpha)} \left[x_i^{(\alpha)} \right]^2 \geq 0 \quad (\text{II.2.48})$$

$$(i=1,2,\dots,n)$$

geldt, overeenkomstig een analoge afleiding als die welke leidde tot (II.2.46), eveneens dat:

$$\sum_{i=1}^n \rho_i = 1 \quad (\text{II.2.49})$$

Dit feit en het feit dat $0 \leq \rho_i \leq 1$, is in de inleiding van paragraaf (II.2) aangehaald, bij de introductie van de entropiefunctie I .

We zullen nu nagaan welke de consequentie is op de autocorrelatiefunctie $K_z(t,t')$, indien $\underline{z}(t)$ eerst in het $\{\psi_i(t)\}$ -coördinaatensysteem en daarna in het Karhunen-Loève $\{\phi_k(t)\}$ -coördinaatensysteem worden beschreven.

In het bijzonder geldt onze belangstelling de relatie tussen de "matrices" $x_{ii'}^{(\alpha)}$ en $z_{kk'}^{(\alpha)}$, beide deeluitmakend van de in (II.2.18) respectievelijk (II.2.40) afgeleide autocorrelatiefunctie. Daartoe herleiden we de uitdrukking (II.2.19),

$$x_{ii'}^{(\alpha)} = \sum_{\alpha} p^{(\alpha)} x_i^{(\alpha)} x_{i'}^{(\alpha)},$$

rekening houdend met (II.2.41), tot:

$$\begin{aligned} & \sum_{\alpha} p^{(\alpha)} \sum_{k=1}^n c_k^{(\alpha)} d_{ki} \sum_{k'=1}^n c_{k'}^{(\alpha)} d_{k'i'} \\ &= \sum_{k=1}^n \sum_{k'=1}^n d_{ki} \sum_{\alpha} p^{(\alpha)} c_k^{(\alpha)} c_{k'}^{(\alpha)} d_{i'k'} \end{aligned}$$

$$= \sum_{k=1}^n \sum_{k'=1}^n d_{ki} \lambda_{k'} \delta_{kk'} d_{i'k'} \quad (\text{II.2.50})$$

In bovenstaande herleiding is gebruik gemaakt van het feit dat:

$$x_i^{(\alpha)} = \sum_{k=1}^n c_k^{(\alpha)} d_{ki} \quad (\text{II.2.51})$$

Dit volgt namelijk uit de in (II.2.39) afgeleide coëfficiënt voor $c_k^{(\alpha)}$:

$$c_k^{(\alpha)} = \sum_{i=1}^n x_i^{(\alpha)} d_{ik}$$

Navermenigvuldiging van linker- en rechterlid met $d_{ki'}$, gevolgd door een sommatie over k resulteert in:

$$\sum_{k=1}^n c_k^{(\alpha)} d_{ki'} = \sum_{k=1}^n \sum_{i=1}^n x_i^{(\alpha)} d_{ik} d_{ki'} = \sum_{i=1}^n x_i^{(\alpha)} \sum_{k=1}^n d_{ik} d_{ki'} \quad (\text{II.2.52})$$

Deze vorm gaat dan, onder gebruikmaking van (II.2.36), over in:

$$\sum_{k=1}^n c_k^{(\alpha)} d_{ki'} = \sum_{i=1}^n x_i^{(\alpha)} \delta_{ii'} = x_i^{(\alpha)}.$$

Uit (II.2.50) volgen, door i gelijk te stellen aan i' , de diagonaalelementen ρ_i van $X_{ii}^{(\alpha)}$,

$$\begin{aligned} \rho_i &= \sum_{k=1}^n (d_{ik})^2 \lambda_k \\ &= \sum_{k=1}^n D_{ik} \lambda_k ; \end{aligned} \quad (\text{II.2.53})$$

waarin

$$D_{ik} = d_{ki} d_{ik} = (d_{ik})^2 \geq 0 \quad (\text{II.2.54})$$

$Z_{kk}^{(\alpha)}$ is, zoals in (II.2.41) is afgeleid een "diagonaal matrix".

De relatie tussen de hoofddiagonaalelementen ρ_i van $X_{ii}^{(\alpha)}$ en de hoofddiagonaalelementen λ_k van $Z_{kk}^{(\alpha)}$, wordt dus gegeven door (II.2.53). Er zij bovendien nog aan herinnerd dat de λ_k 's de eigenwaarden zijn van de integraalvergelijking (II.2.23).

Aangezien de transformatiematrix van het Karhunen-Loève $\{\Phi_k(t)\}$ -coördinaatassenstelsel naar het $\{\Psi_i(t)\}$ -stelsel of omgekeerd orthonormaal is, immers krachtens (II.2.36) geldt:

$$\sum_{k=1}^n d_{ik} d_{ki} = \delta_{ii},$$

$$\sum_{i=1}^n d_{ki} d_{ik} = \delta_{kk},$$

volgt voor de matrix $\|D_{ik}\|$, gezien (II.2.54) dat:

$$\sum_{k=1}^n D_{ik} = \sum_{i=1}^n D_{ik} = 1 \quad (\text{II.2.55})$$

Dat beschrijving van $\underline{z}(t)$ in het Karhunen-Loève coördinaatassenstelsel optimaal is, in die zin dat dan de entropiefunctie I minimaal is, komt neer op het bewijzen van:

$$-\sum_{i=1}^n \rho_i \ln \rho_i \geq -\sum_{k=1}^n \lambda_k \ln \lambda_k \quad (\text{II.2.56})$$

indien gegeven is dat:

$$\lambda_k \geq 0, \quad \sum_{k=1}^n \lambda_k = 1 \quad (\text{II.2.57})$$

$$\rho_i = \sum_{k=1}^n D_{ik} \lambda_k \quad (i=1,2,\dots,n) \quad (\text{II.2.58})$$

waarin:

$$D_{ik} \geq 0 \quad (i,k=1,2,\dots,n)$$

zodanig dat:

$$\sum_{k=1}^n D_{ik} = 1 \quad \text{en} \quad \sum_{i=1}^n D_{ik} = 1 \quad (\text{II.2.59})$$

Daartoe beschouwen we D_{ik} als de kansen op de gebeurtenissen λ_k , ($k=1,2,\dots,n$), van een (nù te introduceren) stochastische grootheid \underline{u} .

Dat wil zeggen:

$$P[\underline{u} = \lambda_k] = D_{ik} \quad (k=1,2,\dots,n) \quad (\text{II.2.60})$$

In wezen zijn hiermee n kansverdelingen gedefiniëerd; voor elke $i=1,2,\dots,n$ namelijk een andere.

Voor de functie $g(\underline{u}) = - \underline{u} \ln \underline{u}$ geldt:

$$g(E \underline{u}) \geq E[g(\underline{u})]^\dagger \quad (\text{II.2.61})$$

Hierin geeft het symbool E de verwachtingswaarde van een stochastische variabele aan.

Aangezien:

$$E \underline{u} = \sum_{k=1}^n D_{ik} \lambda_k$$

volgt uit (II.2.61) dat:

$$- \left\{ \sum_{k=1}^n D_{ik} \lambda_k \right\} \ln \left\{ \sum_{k=1}^n D_{ik} \lambda_k \right\} \geq - \sum_{k=1}^n D_{ik} \left\{ \lambda_k \ln \lambda_k \right\} \\ (i=1, 2, \dots, n)$$

Sommatie van deze uitdrukking over i geeft:

$$- \sum_{i=1}^n \left[\left\{ \sum_{k=1}^n D_{ik} \lambda_k \right\} \ln \left\{ \sum_{k=1}^n D_{ik} \lambda_k \right\} \right] \geq - \sum_{i=1}^n \left[\sum_{k=1}^n D_{ik} \left\{ \lambda_k \ln \lambda_k \right\} \right]$$

Overeenkomstig (II.2.58) en (II.2.59) gaat deze uitdrukking over in:

$$- \sum_{i=1}^n \rho_i \ln \rho_i \geq - \sum_{k=1}^n \lambda_k \ln \lambda_k.$$

† Voor het bewijs hiervan verwijzen wij naar:

Hiermee is dus bewezen dat het Karhunen-Loève coördinaatassenstelsel $\{\phi_k(t)\}$, waarbij de functies $\phi_k(t)$ de eigenfunctie-oplossingen zijn van de integraal vergelijking:

$$\int_a^b K_z(t, t') \phi_k(t') dt' = \lambda_k \phi_k(t)$$

tevens de entropiefunctie I van $\underline{z}(t)$ minimaliseert. Dit betekent, dat het stochastisch proces $\underline{z}(t)$ optimaal -in de zin van minimale entropie- kan worden beschreven in het $\{\phi_k(t)\}$ -stelsel.

HOOFDSTUK III CANONIEKE ONTWIKKELINGEN VAN STOCHASTISCHE PROCESSEN

III.1 Inleiding

Het voor een optimale beschrijving van het stochastische proces $\underline{z}(t)$, noodzakelijkstelsel van coördinaatfuncties $\{\Phi_k(t)\}$, bestaat, zoals in hoofdstuk II werd aangetoond, uit de eigenfunctieoplossingen van de integraal vergelijking (II.2.23), namelijk:

$$\int_a^b K_z(t, t') \Phi_k(t') dt' = \lambda_k \Phi_k(t)$$

In dit hoofdstuk zullen we aantonen dat door toepassing van de methode van de canonieke ontwikkeling van stochastische processen, de oplossing van deze integraalvergelijking tot enkele eenvoudige operaties kan worden teruggebracht. Bovendien bezit deze methode nog het voordeel, dat de voor het stochastische proces $\underline{z}(t)$ kenmerkende eigenschappen, zoals haar trend, periodiciteiten, en andere, afgeleid uit bijvoorbeeld een tijdreeksanalytisch onderzoek, op een zinvolle manier in de coördinaatfuncties $\Phi_k(t)$ kunnen worden verwerkt, en daar zelfs een essentieel onderdeel van vormen.

Een canonieke ontwikkeling van het stochastische proces $\underline{z}(t)$ is gedefiniëerd als elke ontbinding van $\underline{z}(t)$ in de vorm:

$$\underline{z}(t) = \sum_k \underline{x}_k \Phi_k(t), \tag{III.1.1}$$

waarin: $\Phi_k(t)$ niet-stochastische coördinaatfuncties, en
 \underline{x}_k stochastische, niet-gecorrleerde variabelen zijn.

Door de canonieke ontwikkeling (III.1.1) van het stochastische

proces $\underline{z}(t)$, gaat het tijdafhankelijke karakter van $\underline{z}(t)$ over op de niet-stochastische coördinaatfuncties $\Phi_k(t)$, terwijl het stochastische karakter overgaat op de coëfficiënten \underline{x}_k .

De op stochastische processen uit te voeren operaties, behorend bij de oplossing van een differentiaal- of integraal vergelijking, moeten dus, nadat het stochastische proces in canonieke vorm is gebracht, nú uitgevoerd worden op de niet-stochastische coördinaatfuncties $\Phi_k(t)$.

De correlatiefunctie van $\underline{z}(t)$ is per definitie gelijk aan:

$$K_z(t, t') = E [\underline{z}(t) \underline{z}(t')] \quad (\text{III.1.2})$$

Door $\underline{z}(t)$ te ontwikkelen volgens (III.1.1), waarbij zonder verlies aan algemeenheid de verwachtingswaarde van \underline{x}_k gelijk aan nul gesteld kan worden, wordt de volgende uitdrukking voor de correlatiefunctie verkregen:

$$K_z(t, t') = \sum_k D_k \Phi_k(t) \Phi_k(t'), \quad (\text{III.1.3})$$

waarin D_k de variantie van \underline{x}_k voorstelt.

De correlatiefunctie die als kern voorkomt in de integraalvergelijking (II.2.23) is door de ontwikkeling volgens (III.1.1) tot een hanteerbare vorm gebracht. De uitdrukking (III.1.3) zal dan ook een centrale plaats innemen in de in dit hoofdstuk te ontwikkelen procedure ter oplossing van de integraalvergelijking (II.2.23):

III.2 De canonieke ontwikkeling van een stochastisch proces over een rij van discrete punten.

Aan de coëfficiënten \underline{x}_k van de coördinaatfuncties $\Phi_k(t)$ van de canonieke ontwikkeling van het stochastische proces $\underline{z}(t)$, volgens:

$$\underline{z}(t) = \sum_{k=1}^n \underline{x}_k \Phi_k(t) \quad (\text{III.2.1})$$

worden de volgende voorwaarden opgelegd:

$$E[\underline{x}_k] = 0 \quad ; \quad E[\underline{x}_k \underline{x}_j] = 0 \quad \text{voor } k \neq j \quad (\text{III.2.2})$$

(k, j = 1, 2, \dots, n)

Dit betekent dat de covariantie van \underline{x}_k en \underline{x}_j gelijk is aan:

$$\text{Cov}[\underline{x}_k, \underline{x}_j] = E[\underline{x}_k \underline{x}_j] - E[\underline{x}_k] E[\underline{x}_j] = 0 \quad (\text{III.2.3})$$

terwijl voor de variantie van \underline{x}_j geldt:

$$\text{Var}[\underline{x}_j] = E[\underline{x}_j^2] - [E \underline{x}_j]^2 = D_j \quad (\text{III.2.4})$$

Gebruikmakend van de ontwikkeling van $\underline{z}(t)$ volgens (III.2.1) en rekening houdend met de voorwaarden (III.2.2) en (III.2.3) wordt:

$$E[\underline{z}(t) \underline{x}_j] = \sum_{k=1}^n \Phi_k(t) E[\underline{x}_k \underline{x}_j] = D_j \Phi_j(t),$$

waaruit volgt:

$$\Phi_j(t) = \frac{1}{D_j} E[\underline{z}(t) \underline{x}_j] \quad \text{voor } j=1, 2, \dots, n \quad (\text{III.2.5})$$

Deze uitdrukking laat zien, dat, wil het proces $\underline{z}(t)$ ontwikkeld kunnen worden volgens (III.2.1), dat wil zeggen, uitgedrukt kunnen

worden in termen van de functies $\Phi_k(t)$, het dan noodzakelijk is, dat alle coëfficiënten \underline{x}_j gecorreleerd zijn met $\underline{z}(t)$. Daarom beschouwen we voor \underline{x}_j , de volgende lineaire combinatie van de stochastische grootheden $\underline{z}(t_h)$:

$$\underline{x}_j = \sum_{h=1}^N a_{jh} \underline{z}(t_h) \quad (j=1,2,\dots,n) \quad (\text{III.2.6})$$

Hierin is $\underline{z}(t_h)$ het stochastische proces $\underline{z}(t)$ op het tijdstip $t = t_h$; de coëfficiënt a_{jh} is voorlopig nog willekeurig, waarbij $h = 1,2,3,\dots,N$.

Substitutie van (III.2.6) in (III.2.5) geeft, rekening houdend met (III.1.2):

$$\begin{aligned} \Phi_j(t) &= \frac{1}{D_j} \sum_{h=1}^N a_{jh} E [\underline{z}(t) \underline{z}(t_h)] \\ &= \frac{1}{D_j} \sum_{h=1}^N a_{jh} K_z(t, t_h) \end{aligned} \quad (\text{III.2.7})$$

Uitgaande van (III.2.6) kunnen de varianties en covarianties van \underline{x}_j en \underline{x}_k worden bepaald.

Deze worden, overeenkomstig (III.2.4) en (III.2.3), rekening houdend met (III.1.2):

$$\begin{aligned} \text{Var } \underline{x}_j = D_j &= \sum_{h=1}^N \sum_{i=1}^N a_{jh} a_{ji} E [\underline{z}(t_h) \underline{z}(t_i)] \\ &= \sum_{h=1}^N \sum_{i=1}^N a_{jh} a_{ji} K_z(t_h, t_i) \end{aligned} \quad (\text{III.2.8})$$

$$\begin{aligned} \text{Cov} [\underline{x}_j \underline{x}_k] &= \sum_{h=1}^N \sum_{i=1}^N a_{jh} a_{ki} E [\underline{z}(t_h) \underline{z}(t_i)] \\ &= \sum_{h=1}^N \sum_{i=1}^N a_{jh} a_{ki} K_z(t_h, t_i) \end{aligned} \quad (\text{III.2.9})$$

voor $j \neq k$

Aangezien de covariantie tussen \underline{x}_j en \underline{x}_k , overeenkomstig de voorwaarde (III.2.3), gelijk is aan nul, volgt uit (III.2.9) dat de coëfficiënten a_{jh} van de lineaire combinatie (III.2.6), moeten voldoen aan de volgende voorwaarden:

$$\sum_{h=1}^N \sum_{i=1}^N a_{jh} a_{ki} K_z(t_h, t_i) = 0 \quad \text{voor } j \neq k \quad (\text{III.2.10})$$

($j, k = 1, 2, \dots, n$)

Door substitutie van (III.2.7), voor $t = t_i$, in (III.2.8) respectievelijk (III.2.9) zijn variantie en covariantie ook te schrijven als:

$$\text{Var } \underline{x}_j = D_j = D_j \sum_{i=1}^N a_{ji} \Phi_j(t_i) \quad (\text{III.2.11})$$

$$\text{Cov} [\underline{x}_j \underline{x}_k] = D_j \sum_{i=1}^N a_{ki} \Phi_j(t_i) \quad (\text{III.2.12})$$

Dit betekent, gezien de voorwaarden (III.2.3), dat:

$$\begin{aligned} \sum_{i=1}^N a_{ki} \Phi_j(t_i) &= 1 \quad \text{voor } j = k \\ &= 0 \quad \text{voor } j \neq k \end{aligned} \quad (\text{III.2.13})$$

($j, k = 1, 2, \dots, n$)

Aangezien het aantal vergelijkingen in (III.2.10) altijd kleiner is dan het aantal onbekende coëfficiënten a_{jh} uit de lineaire combinatie (III.2.6), kunnen deze onbekenden op een oneindig groot aantal manieren worden bepaald.

Eén van deze mogelijkheden is de volgende:

Bepaal voor een gegeven realisatie $z(t)$ van $\underline{z}(t)$ de waarde x_1 van \underline{x}_1 door aan alle coëfficiënten a_{1h} willekeurige waarden toe te kennen.

De onbekenden a_{2h} echter moeten zodanig bepaald worden dat \underline{x}_1 en \underline{x}_2 ongecorreleerd zijn; dit betekent dat aan alle onbekenden a_{2h} , op één na, willekeurige waarden kunnen worden toegekend. Immers, uit (III.2.10) volgt, dat, indien alle coëfficiënten a_{1h} zijn vastgelegd, evenals alle coëfficiënten a_{2h} op de laatste na, deze laatste zodanig gekozen moet worden dat aan (III.2.10) is voldaan. Achtereenvolgens kunnen op deze wijze de realisaties x_3, x_4, \dots van $\underline{x}_3, \underline{x}_4, \dots$ worden bepaald.

De onbekende coëfficiënten a_{nh} van x_n dienen dus zodanig bepaald te worden dat \underline{x}_n niet gecorreleerd is met $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_{n-1}$; dit betekent toekenning van willekeurige waarden aan alle coëfficiënten a_{nh} met uitzondering van $(n-1)$ daarvan, omdat in dit geval $(n-1)$ vergelijkingen als in (III.2.10) voorkomen.

We kunnen de onbekende coëfficiënten a_{jh} bijvoorbeeld ook op de volgende wijze specificeren:

$$\begin{aligned} a_{jj} &= 1 \\ a_{jh} &= 0 \quad \text{voor } h > j \quad \begin{matrix} (j=1,2,\dots,n) \\ (h=1,2,\dots,N) \end{matrix} \quad \text{(III.2.14)} \end{aligned}$$

Substitutie van $j = 1$ in de uitdrukkingen (III.2.6), (III.2.7) en (III.2.8) geeft dan, gebruikmakend van (III.2.14),

$$\begin{aligned}
 \underline{x}_1 &= \underline{z}(t_1) \\
 \Phi_1(t) &= \frac{1}{D_1} K_z(t, t_1) \\
 D_1 &= K_z(t_1, t_1)
 \end{aligned} \tag{III.2.15}$$

Voor $j = 2, 3, \dots$, gaan de uitdrukkingen (III.2.6), (III.2.7) en (III.2.8) achtereenvolgens over in:

$$\begin{aligned}
 \underline{x}_j &= \sum_{h=1}^{j-1} a_{jh} \underline{z}(t_h) + \underline{z}(t_j) \\
 \Phi_j(t) &= \frac{1}{D_j} \left[\sum_{h=1}^{j-1} a_{jh} K_z(t, t_h) + K_z(t, t_j) \right] \\
 D_j &= \sum_{h=1}^{j-1} \sum_{i=1}^{j-1} a_{jh} a_{ji} K_z(t_h, t_i) + \sum_{h=1}^{j-1} a_{jh} K_z(t_h, t_j) \\
 &\quad + \sum_{i=1}^{j-1} a_{ji} K_z(t_i, t_j) + K_z(t_j, t_j)
 \end{aligned} \tag{III.2.16}$$

($j=1, 2, \dots, n$)

Als slot van deze paragraaf een tweetal interessante eigenschappen van de canonieke ontwikkeling van stochastische processen.

De eerste daarvan luidt als volgt:

Worden de coëfficiënten \underline{x}_j gespecificeerd volgens (III.2.6) en de coördinaatfuncties $\Phi_j(t)$ volgens (III.2.7), dan wordt door de canonieke ontwikkeling (III.2.1), het stochastisch proces $\underline{z}(t)$ in de punten $t = t_i$ exact weergegeven.

Voor het bewijs hiervan substitueren we (III.2.6) in (III.2.1);
dit geeft voor $t = t_i$:

$$\begin{aligned} \underline{z}(t_i) &= \sum_{j=1}^n \underline{x}_j \Phi_j(t_i) = \sum_{j=1}^n \sum_{h=1}^N a_{jh} \underline{z}(t_h) \Phi_j(t_i) \\ &= \sum_{h=1}^N \underline{z}(t_h) \sum_{j=1}^n a_{jh} \Phi_j(t_i) = \sum_{h=1}^N \underline{z}(t_h) b_{hi} \quad (\text{III.2.17}) \end{aligned}$$

waarin:

$$b_{hi} = \sum_{j=1}^n a_{jh} \Phi_j(t_i) \quad (\text{III.2.18})$$

Indien nu bewezen kan worden dat:

$$\begin{aligned} &= 1 \quad \text{voor} \quad h \neq i \\ b_{hi} &= \delta_{hi} \quad (\text{III.2.19}) \\ &= 0 \quad \text{voor} \quad h \neq i \end{aligned}$$

dan volgt uit (III.2.17) het gestelde.

Om (III.2.19) te bewijzen gaan we uit van (III.2.13):

$$\sum_{i=1}^N a_{ki} \Phi_j(t_i) = \delta_{jk} \quad (j, k=1, 2, \dots, n)$$

Uit (III.2.6) blijkt dat de index k van a_{ki} betrekking heeft op de coëfficiënt \underline{x}_k van de coördinaatfunctie $\Phi_k(t)$, evenals de index j van a_{jh} , afkomstig van de uitdrukking (III.2.18), behoort bij de coëfficiënt \underline{x}_j van de coördinaatfunctie $\Phi_j(t)$. Laten we er van

uitgaan dat de ontwikkeling van $\underline{z}(t)$ volgens (III.1.1) n functies $\Phi(t)$ bevat. Dit wil dus zeggen, dat $j, k = 1, 2, \dots, n$. De index i daarentegen heeft betrekking op de tijdstippen $t = t_i$ -waarbij $i = 1, 2, \dots, N$ - waarop het stochastische proces $\underline{z}(t)$ wordt waargenomen.

De uitdrukkingen (III.2.13) kunnen worden samengevat tot het volgende product van 2 matrices:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1i} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2i} & \dots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{k1} & a_{k2} & \dots & a_{ki} & \dots & a_{kN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{ni} & \dots & a_{nN} \end{pmatrix} \begin{pmatrix} \Phi_1(t_1) & \Phi_2(t_1) & \dots & \Phi_j(t_1) & \dots & \Phi_n(t_1) \\ \Phi_1(t_2) & \Phi_2(t_2) & \dots & \Phi_j(t_2) & \dots & \Phi_n(t_2) \\ \vdots & \vdots & & \vdots & & \vdots \\ \Phi_1(t_i) & \Phi_2(t_i) & \dots & \Phi_j(t_i) & \dots & \Phi_n(t_i) \\ \vdots & \vdots & & \vdots & & \vdots \\ \Phi_1(t_N) & \Phi_2(t_N) & \dots & \Phi_j(t_N) & \dots & \Phi_n(t_N) \end{pmatrix} = I,$$

waarin I de eenheidsmatrix voorstelt.

Dit betekent evenwel, dat (III.2.13) ook kan worden geschreven als:

$$\begin{pmatrix} \Phi_1(t_1) & \Phi_2(t_1) & \dots & \Phi_j(t_1) & \dots & \Phi_n(t_1) \\ \Phi_1(t_2) & \Phi_2(t_1) & \dots & \Phi_j(t_2) & \dots & \Phi_n(t_2) \\ \vdots & \vdots & & \vdots & & \vdots \\ \Phi_1(t_i) & \Phi_2(t_i) & \dots & \Phi_j(t_i) & \dots & \Phi_n(t_i) \\ \vdots & \vdots & & \vdots & & \vdots \\ \Phi_1(t_N) & \Phi_2(t_N) & \dots & \Phi_j(t_N) & \dots & \Phi_n(t_N) \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1i} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2i} & \dots & a_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{k1} & a_{k2} & \dots & a_{ki} & \dots & a_{kN} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{ni} & \dots & a_{nN} \end{pmatrix} = I,$$

(III.2.20)

Immers, indien het product van 2 matrices de eenheidsmatrix oplevert, met andere woorden:

$$A \Phi = I$$

dan is ook:

$$\Phi A = I$$

Het product van de matrices A en Φ uit (III.2.20) kan nu weer worden geschreven in de vorm van een vergelijkingenstelsel, namelijk volgens:

$$\sum_{j=1}^n \Phi_j(t_i) a_{jh} = \sum_{j=1}^n a_{jh} \Phi_j(t_i) = \delta_{hi} \quad (\text{III.2.21})$$

waarmee (III.2.19) eveneens is bewezen.

De tweede, voor de canonieke ontwikkelingen van stochastische processen, interessante eigenschap houdt het volgende in:

De canonieke ontwikkeling van $\underline{z}(t)$ volgens (III.2.1), waarbij de coördinaatfuncties $\Phi_k(t)$ en de coëfficiënten \underline{x}_k verbonden zijn via de relatie (III.2.5), waarborgt voor elke t de beste benadering van $\underline{z}(t)$; in die zin, dat de mathematische verwachting van het kwadraat van de residuele term, behorend bij de $\{\Phi_k(t)\}$ -ontwikkeling volgens (III.2.1), altijd kleiner zal zijn, dan die welke behoort bij de ontwikkeling van $\underline{z}(t)$ in een willekeurig ander stelsel van coördinaatfuncties, waarbij in beide gevallen de ontwikkeling met behulp van een even groot aantal functies dient te geschieden. Het $\{\Phi_k(t)\}$ -stelsel van coördinaatfuncties is dus ook in deze zin optimaal.

Wordt $\underline{z}(t)$ bijvoorbeeld ontwikkeld in een willekeurig $\{\Psi_j(t)\}$ -stelsel, waarbij we ons beperken tot n functies $\Psi_j(t)$, volgens:

$$\underline{z}(t) = \sum_{j=1}^n \underline{x}_j' \Psi_j(t),$$

dan kan het residu weergegeven worden door:

$$\underline{R}_n(t) = \underline{z}(t) - \sum_{j=1}^n \underline{x}_j' \Psi_j(t)$$

In de $\{\Phi_j(t)\}$ -ontwikkeling, waarbij $\Phi_j(t)$ bepaald wordt volgens (III.2.5), is dit residu, overeenkomstig de zojuist beschreven eerste eigenschap van de canonieke ontwikkeling, voor elke $t = t_i$, gelijk aan nul.

De mathematische verwachting van het kwadraat van $R_n(t)$, wordt, rekening houdend met (III.2.3), (III.2.4), en (III.2.5):

$$\begin{aligned} E [\underline{R}_n(t)]^2 &= E \left[\underline{z}(t) - \sum_{j=1}^n \underline{x}_j' \Psi_j(t) \right]^2 \\ &= E \left[\underline{z}(t) - \sum_{j=1}^n \underline{x}_j' \left\{ \Psi_j(t) - \Phi_j(t) + \Phi_j(t) \right\} \right]^2 \\ &= E \left[\underline{z}(t) - \sum_{j=1}^n \underline{x}_j' \left\{ \Psi_j(t) - \Phi_j(t) \right\} - \sum_{j=1}^n \underline{x}_j' \Phi_j(t) \right]^2 \\ &= E \left[\underline{z}(t) \right]^2 + \sum_{j=1}^n D_j \left\{ \Psi_j(t) - \Phi_j(t) \right\}^2 + \sum_{j=1}^n D_j \left\{ \Phi_j(t) \right\}^2 \\ &\quad - 2 E \left[\underline{z}(t) \cdot \sum_{j=1}^n \underline{x}_j' \left\{ \Psi_j(t) - \Phi_j(t) \right\} \right] \\ &\quad - 2 E \left[\underline{z}(t) \cdot \sum_{j=1}^n \underline{x}_j' \Phi_j(t) \right] \end{aligned}$$

$$\begin{aligned}
 & + 2 E \left[\sum_{j=1}^n \underline{x}_j' \Phi_j(t) \cdot \sum_{j=1}^n \underline{x}_j' \left\{ \Psi_j(t) - \Phi_j(t) \right\} \right] \\
 & = E \left[\underline{z}(t) \right]^2 + \sum_{j=1}^n D_j \left\{ \Psi_j(t) - \Phi_j(t) \right\}^2 + \sum_{j=1}^n D_j \left\{ \Phi_j(t) \right\}^2 \\
 & - 2 \sum_{j=1}^n D_j \Phi_j(t) \left\{ \Psi_j(t) - \Phi_j(t) \right\} \\
 & - 2 \sum_{j=1}^n D_j \left\{ \Phi_j(t) \right\}^2 + 2 \sum_{j=1}^n D_j \Phi_j(t) \left\{ \Psi_j(t) - \Phi_j(t) \right\} \\
 & = E \left[\underline{z}(t) \right]^2 - \sum_{j=1}^n D_j \left\{ \Phi_j(t) \right\}^2 + \sum_{j=1}^n D_j \left\{ \Psi_j(t) - \Phi_j(t) \right\}^2
 \end{aligned}$$

Uit deze laatste uitdrukking valt onmiddellijk te zien dat de verwachting van het kwadraat van het residu $R_n(t)$, voor elke t minimaal is, indien:

$$\Psi_j(t) \equiv \Phi_j(t)$$

voor $j = 1, 2, \dots, n$

III.3 De canonieke ontwikkeling van een stochastisch proces over een bepaald interval.

In plaats van de uitdrukking (III.2.6) voor de coëfficiënten \underline{x}_j van de canonieke ontwikkeling (III.2.1) van het stochastisch proces $\underline{z}(t)$, beschouwen nu:

$$\underline{x}_j = \int_a^b a_j(t) \underline{z}(t) dt \quad a \leq t \leq b \quad (\text{III.3.1})$$

$$(j = 1, 2, \dots, n)$$

De enige restrictie die aan de functies $a_k(t)$ en $a_j(t)$ wordt opgelegd, volgt uit de voorwaarde dat \underline{x}_k en \underline{x}_j ongecorreleerd moeten zijn. Dit betekent, uitgaand van (III.3.1):

$$E [\underline{x}_k \underline{x}_j] = \int_a^b \int_a^b a_k(t) a_j(t') K_z(t, t') dt dt'$$

$$= 0 \quad \text{voor } k \neq j \quad (\text{III.3.2})$$

Stellen we in (III.3.2), k gelijk aan j , dan wordt de variantie D_j van \underline{x}_j verkregen:

$$\text{Var } \underline{x}_j = D_j = \int_a^b \int_a^b a_j(t) a_j(t') K_z(t, t') dt dt' \quad (\text{III.3.3})$$

Om nu de coördinaatfunctie $\Phi_j(t)$ te vinden substitueren we (III.3.1) in (III.2.5),

$$\Phi_j(t) = \frac{1}{D_j} E [\underline{z}(t) \underline{x}_j]$$

Dit levert:

$$\begin{aligned}
 \Phi_j(t) &= \frac{1}{D_j} E \left[\underline{z}(t) \int_a^b a_j(t') \underline{z}(t') dt' \right] \\
 &= \frac{1}{D_j} \int_a^b a_j(t') E \left[\underline{z}(t) \underline{z}(t') \right] dt' \\
 &= \frac{1}{D_j} \int_a^b a_j(t') K_z(t, t') dt' \quad (\text{III.3.4})
 \end{aligned}$$

Substitutie van (III.3.4) in (III.3.2) geeft:

$$\begin{aligned}
 E [\underline{x}_k \underline{x}_j] &= \int_a^b \int_a^b a_k(t) a_j(t') K_z(t, t') dt dt' \\
 &= \int_a^b a_k(t) \left[\int_a^b a_j(t') K_z(t, t') dt' \right] dt \\
 &= D_j \int_a^b a_k(t) \Phi_j(t) dt \quad (\text{III.3.5})
 \end{aligned}$$

Uit (III.3.2), (III.3.3) en (III.3.5) volgen dan de relaties welke moeten bestaan tussen de functies $a_k(t)$ en de coördinaatfuncties $\Phi_j(t)$, namelijk:

$$\int_a^b a_k(t) \phi_j(t) dt = \delta_{kj} \quad \begin{aligned} &= 1 \text{ voor } k = j \\ &= 0 \text{ voor } k \neq j \end{aligned} \quad (\text{III.3.6})$$

Deze uitdrukking vertoont, zoals te verwachten, grote overeenkomst met (III.2.13); is daar het continue analogon van.

Uit (III.3.2) is tevens, rekening houdend met (III.3.4), de volgende relatie af te leiden:

$$\begin{aligned} E[\underline{x}_k \underline{x}_j] &= \int_a^b \int_a^b a_k(t) a_j(t') K_z(t, t') dt dt' \\ &= \int_a^b a_j(t') dt' \left[\int_a^b a_k(t) K_z(t', t) dt \right] \\ &= D_k \int_a^b a_j(t') \phi_k(t') dt' \end{aligned} \quad (\text{III.3.7})$$

Gecombineerd met (III.3.5) levert dit:

$$D_j \int_a^b a_k(t) \phi_j(t) dt = D_k \int_a^b a_j(t) \phi_k(t) dt \quad (\text{III.3.8})$$

De uitdrukkingen (III.3.8) worden de dubbel-orthogonaliteits voorwaarden van de functies $a_k(t)$ en $\phi_j(t)$ genoemd. Indien dus de functie $a_k(t)$ orthogonaal is met de functie $\phi_j(t)$, dan is tevens de functie $a_j(t)$ orthogonaal met de functie $\phi_k(t)$.

Deze eigenschap stelt ons in staat om achtereenvolgens de paren van functies $a_j(t)$ en $\phi_j(t)$ af te leiden, welke voldoen aan (III.3.4) en (III.3.6).

Zou namelijk $a_n(t)$ orthogonaal zijn met $\phi_1(t), \phi_2(t), \dots, \phi_{n-1}(t)$, dan is volgens (III.3.8) de functie $\phi_n(t)$ orthogonaal met $a_1(t), a_2(t), \dots, a_{n-1}(t)$.

Dit betekent dan, dat aan (III.3.4) en (III.3.6) is voldaan voor alle functies $a_k(t)$ en $\phi_j(t)$, waarbij $k, j = 1, 2, \dots, n$.

Elk zo'n stelsel van functies $\{a_k(t)\}$ en $\{\phi_j(t)\}$ levert een canonieke ontwikkeling van het stochastisch proces $\underline{z}(t)$.

Om een stelsel van functies $\{a_k(t)\}$ en $\{\phi_j(t)\}$ te vinden, dat voldoet aan (III.3.4) en (III.3.6) stellen we de volgende methode voor, welke ons successievelijk alle functies $a_j(t)$ en $\phi_j(t)$ zal opleveren. Deze methode gaat uit van de in eerste instantie volkomen willekeurige hulpfuncties $\eta_1(t), \eta_2(t), \dots, \eta_n(t)$.

Stel nu:

$$a_1(t) = \eta_1(t)$$

Dan zijn de variantie D_1 en de eerste coördinaatfunctie $\phi_1(t)$ te bepalen met behulp van (III.3.3) respectievelijk (III.3.4).

De functie $a_2(t)$ schrijven we als:

$$a_2(t) = p_{21} a_1(t) + \eta_2(t), \quad (\text{III.3.9})$$

waarbij de constante p_{21} zodanig dient te worden bepaald, dat de functies $a_2(t)$ en $\phi_1(t)$ voldoen aan (III.3.6), dat wil zeggen:

$$\int_a^b a_2(t) \phi_1(t) dt = p_{21} \int_a^b a_1(t) \phi_1(t) dt + \int_a^b \eta_2(t) \phi_1(t) dt$$

$$= p_{21} + \int_a^b \eta_2(t) \Phi_1(t) dt$$

$$= 0$$

Hieruit volgt:

$$p_{21} = - \int_a^b \eta_2(t) \Phi_1(t) dt, \quad (\text{III.3.10})$$

waarna de variantie D_2 en de tweede coördinaatfunctie $\Phi_2(t)$ bepaald kunnen worden aan de hand van (III.3.3) en (III.3.4).

De op deze wijze, achtereenvolgens bepaalde functies $a_j(t)$ en $\Phi_j(t)$ - waarbij $j=1,2,\dots,(n-1)$ - vormen dan het uitgangspunt voor de functie $a_n(t)$, die we gelijk stellen aan:

$$a_n(t) = \sum_{j=1}^{n-1} p_{nj} a_j(t) + \eta_n(t) \quad (\text{III.3.11})$$

De coëfficiënten p_{nj} dienen weer zodanig bepaald te worden, dat de functie $a_n(t)$ orthogonaal is met de functies $\Phi_1(t), \Phi_2(t), \dots, \Phi_k(t), \dots, \Phi_{n-1}(t)$.

Dit betekent, rekening houdend met (III.3.6) dat:

$$\int_a^b a_n(t) \Phi_j(t) dt = \sum_{k=1}^{n-1} p_{nk} \int_a^b a_k(t) \Phi_j(t) dt + \int_a^b \eta_n(t) \Phi_j(t) dt$$

$$= p_{nj} + \int_a^b \eta_n(t) \Phi_j(t) dt$$

$$= 0 \quad j = 1, 2, \dots, (n-1).$$

Hieruit volgt:

$$p_{nj} = - \int_a^b \eta_n(t) \Phi_j(t) dt \quad j = 1, 2, \dots, (n-1) \quad (\text{III.3.12})$$

Door de uitdrukkingen (III.3.11) en (III.3.12) wordt de functie $a_n(t)$ volkomen bepaald; de volgende stap bestaat uit het bepalen van de variantie D_n volgens (III.3.3) en de functie $\Phi_n(t)$ volgens (III.3.4). Krachtens (III.3.8) is de functie $\Phi_n(t)$ orthogonaal met de in eerdere phazen bepaalde functies $a_1(t)$, $a_2(t)$, ..., $a_{n-1}(t)$. Aangezien het stelsel van hulpfuncties $\{\eta_j(t)\}$ op een oneindig groot aantal manieren kan worden gekozen, bestaat er een daarmee overeenkomend aantal stelsels van paren van functies $\{a_j(t)\}$ en $\{\Phi_j(t)\}$.

III.4 De Karhunen-Loève ontwikkeling als een bijzonder geval van de canonieke ontwikkeling.

In hoofdstuk II is de integraal vergelijking:

$$\int_a^b K_z(t, t') \Phi_k(t') dt' = \lambda_k \Phi_k(t) \quad (\text{III.4.1})$$

geïntroduceerd, waarvan het stelsel van eigenfunctie oplossingen $\{\Phi_k(t)\}$ een optimale beschrijving van het stochastische proces $\underline{z}(t)$ blijkt te bewerkstelligen. Deze eigenfuncties hebben, overeenkomstig (II.2.35) nog de eigenschap dat:

$$\int_a^b \Phi_k(t) \Phi_{k'}(t) dt = \delta_{kk'}, \quad (\text{III.4.2})$$

Indien we deze uitdrukkingen, waarin de voorwaarden voor een optimale (in de zin van minimale entropie) ontwikkeling liggen besloten, vergelijken met de uitdrukking (III.3.4):

$$\Phi_k(t) = \frac{1}{D_k} \int_a^b a_k(t') K_z(t, t') dt' ,$$

met andere woorden,

$$D_k \Phi_k(t) = \int_a^b a_k(t') K_z(t, t') dt' \quad (\text{III.4.3})$$

respectievelijk met de uitdrukking (III.3.6):

$$\int_a^b a_k(t) \Phi_{k'}(t) dt = \delta_{kk'} , \quad (\text{III.4.4})$$

dan blijkt de Karhunen-Loève ontwikkeling niets anders te zijn dan een bijzonder geval van de canonieke ontwikkeling, waarbij dan:

$$\Phi_k(t) \equiv a_k(t) \quad (\text{III.4.5})$$

$$\lambda_k \equiv D_k \quad (\text{III.4.6})$$

Het stelsel van coördinaatfuncties $\{\Phi_k(t)\}$ is dus in twee opzichten optimaal.

In de eerste plaats minimaliseert dit stelsel, volgens (II.2.56), de entropie functie.

In de tweede plaats minimaliseert dit stelsel de verwachting van

het kwadraat van de residuele term van $\underline{z}(t)$ in de ontwikkeling volgens (III.1.1)

De uitdrukking (III.4.5) betekent, rekening houdend met (III.3.11), waarin de functie $a_k(t)$ werd afgeleid als lineaire combinatie van de hulpfuncties $\eta_1(t)$, $\eta_2(t)$, ..., dat de k -de coördinaatfunctie van het optimale Karhunen-Loève coördinaatassenstelsel $\{\Phi_k(t)\}$ te schrijven is als:

$$\Phi_k(t) = p_{k1} \eta_1(t) + p_{k2} \eta_2(t) + \dots + p_{kj} \eta_j(t) + \dots + \eta_k(t) \quad (\text{III.4.7})$$

waarbij de coëfficiënten p_{kj} , overeenkomstig (III.3.12), bepaald moeten worden aan de hand van:

$$p_{kj} = - \int_a^b \eta_k(t) \Phi_j(t) dt \quad j = 1, 2, \dots, (k-1)$$

$$(k = 1, 2, \dots, n)$$

Indien in het volgens:

$$\underline{z}(t) = \sum_k \underline{x}_k \Phi_k(t)$$

te ontwikkelen stochastische proces $\underline{z}(t)$, de coördinaatfuncties $\Phi_k(t)$, worden gesubstitueerd, welke volgens (III.4.7) zijn uitgedrukt in het stelsel van hulpfuncties $\{\eta_j(t)\}$, dan krijgen we:

$$\underline{z}(t) = [\underline{x}_1 + \underline{x}_2 p_{21} + \underline{x}_3 p_{31} + \dots] \eta_1(t)$$

$$\begin{aligned}
 & + [\underline{x}_2 + \underline{x}_3 p_{32} + \underline{x}_4 p_{42} + \dots] \eta_2(t) \\
 & + [\underline{x}_3 + \underline{x}_4 p_{43} + \underline{x}_5 p_{53} + \dots] \eta_3(t) \\
 & + \dots \\
 & + \dots \\
 & = \underline{x}_1^* \eta_1(t) + \underline{x}_2^* \eta_2(t) + \dots \underline{x}_3^* \eta_3(t) + \dots \quad (\text{III.4.8})
 \end{aligned}$$

Het stochastisch proces $\underline{z}(t)$ is hiermee ontwikkeld in termen van de functies $\eta_1(t)$, $\eta_2(t)$, $\eta_3(t)$,...

Deze hulpfuncties zijn in eerste instantie volkomen willekeurig verondersteld.

Onze uiteindelijke doelstelling betrof een onderzoek naar het tijdafhankelijke karakter van de met behulp van de methode van de componentenanalyse verkregen componenten, en wel door haar tijdafhanke-lijkheid te beschouwen in relatie tot die van de variabelen

$\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ welke het totale model bepalen. Binnen het model dat door de component zelf wordt gevormd, kan echter tegelijkertijd de onderlinge tijdafhanke-lijkheid van de afzonderlijke variabelen

$\underline{z}_1, \underline{z}_2, \dots, \underline{z}_\ell, \dots, \underline{z}_p$ worden bestudeerd.

Daartoe dient elke variabele \underline{z}_ℓ , -waarbij $\ell=1,2,\dots,p$ -, ontwikkeld te worden volgens (III.4.8) in termen van de voor deze variabele karakteristieke eigenschappen, waarna de dan verkregen ontwikkeling dient te worden gesubstitueerd in de volgende uitdrukking voor de component \underline{f}_i :

$$\underline{f}_i = b_{i1} \underline{z}_1 + b_{i2} \underline{z}_2 + \dots + b_{i\ell} \underline{z}_\ell + \dots + b_{ip} \underline{z}_p.$$

Deze wijze van benaderen maakt de voor economische tijdreeksen, zo weinig succesvolle kruisspectraalanalyse, geheel overbodig.

Worden nu in de functies $\eta_1(t)$, $\eta_2(t)$, $\eta_3(t)$... de uit tijdreeks-analytisch onderzoek gevonden eigenschappen van $\underline{z}_\ell(t)$ verwerkt, dan is met behulp van (III.4.8) bereikt, dat het proces $\underline{z}_\ell(t)$ optimaal beschreven is in termen van functies, die de voor $\underline{z}_\ell(t)$ kenmerkende eigenschappen bevatten.

Zo kan bijvoorbeeld de functie $\eta_1(t)$ de trend van $\underline{z}_\ell(t)$ weergeven, de functie $\eta_2(t)$ een bepaalde periode bevatten, enzovoort.

APPENDIX

De in paragraaf I.3 op bladzijde 30 gebruikte eigenschap met betrekking tot de voorwaardelijke varianties,

$\sigma_{2.1}^2, \sigma_{3.12}^2, \sigma_{4.123}^2, \dots, \sigma_{p.12\dots(p-1)}^2$ komt neer op het bewijzen van:

$$\prod_{i=1}^p \sigma_i^2 |R| = \sigma_1^2 \cdot \sigma_{2.1}^2 \sigma_{3.12}^2 \dots \sigma_{p.12\dots(p-1)}^2 \quad (A1)$$

Het linkerlid hiervan is in wezen de determinant van de variantie-covariantiematrix $\sum_{(pp)}$ van de variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$.

Voor het bewijs van (A1) zullen we gebruik maken van:

1. een eigenschap van gepartitioneerde matrices
2. een eigenschap van de p-dimensionale normale verdeling van $\underline{z} = (\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p)'$,

waarbij:

$$E \underline{z} = E \begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \vdots \\ \underline{z}_p \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \mu \quad \text{en} \quad \text{Var } \underline{z} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & & & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix} = \sum_{(pp)}$$

Samengevat: $\underline{z} \sim N_p(\mu, \sum_{pp})$.

In (I.3) is μ dus de nulvector.

Eigenschap 1

Indien de matrix X te partitioneren is volgens:

$$X = \begin{pmatrix} X^{11} & X^{12} \\ X^{21} & X^{22} \end{pmatrix}$$

waarin de matrices X^{11} en X^{22} beide vierkant zijn, dan is, indien X^{22} regulier is, de determinant van X gelijk aan:

$$\begin{aligned} |X| &= \begin{vmatrix} X^{11} & X^{12} \\ X^{21} & X^{22} \end{vmatrix} = \begin{vmatrix} X^{11} & X^{12} \\ X^{21} & X^{22} \end{vmatrix} \begin{vmatrix} I & 0 \\ -X^{22^{-1}} & I \end{vmatrix} \\ &= \begin{vmatrix} X^{11.2} & X^{12} \\ 0 & X^{22} \end{vmatrix} = |X^{11.2}| |X^{22}| \end{aligned} \quad (A2)$$

Hierin is:

$$X^{11.2} = X^{11} - X^{12} X^{22^{-1}} X^{21} \quad (A3)$$

Wordt nu de determinant X^{11} regulier verondersteld, dan is analoog aan (A2) de determinant van X te schrijven als:

$$|X| = |X^{22.1}| |X^{11}|, \quad (A4)$$

waarin:

$$X^{22.1} = X^{22} - X^{21} X^{11}{}^{-1} X^{12} \quad (A5)$$

Eigenschap 2

We gaan uit van de p variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$ welke p -dimensionaal normaal verdeeld zijn, dat wil zeggen:

$$\underline{z} \sim N_p(\mu, \Sigma_{(pp)}).$$

Splits de p variabelen in twee groepen $\underline{z}^{(1)}$ en $\underline{z}^{(2)}$ van p_1 respectievelijk p_2 elementen.

Dit betekent:

$$\underline{z}^{(1)} = \begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \vdots \\ \underline{z}_{p_1} \end{pmatrix} \quad \underline{z}^{(2)} = \begin{pmatrix} \underline{z}_{p_1+1} \\ \underline{z}_{p_1+2} \\ \vdots \\ \underline{z}_p \end{pmatrix}$$

zodat:

$$\underline{z} = \begin{pmatrix} \underline{z}^{(1)} \\ \underline{z}^{(2)} \end{pmatrix} \quad \mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$$

$$\sum_{(pp)} = \begin{pmatrix} \sum_{(p_1 p_1)}^{11} & \sum_{(p_1 p_2)}^{12} \\ \sum_{(p_2 p_1)}^{21} & \sum_{(p_2 p_2)}^{22} \end{pmatrix}$$

Hierin is dus:

$$\begin{aligned} \sum_{(p_1 p_1)}^{11} &= \text{Var } \underline{z}^{(1)} & ; & \quad \sum_{(p_2 p_2)}^{22} = \text{Var } \underline{z}^{(2)} \\ \sum_{(p_1 p_2)}^{12} &= \text{Cov } (\underline{z}^{(1)}, \underline{z}^{(2)}) & ; & \quad \sum_{(p_2 p_1)}^{21} = \text{Cov } (\underline{z}^{(2)}, \underline{z}^{(1)}) \end{aligned}$$

De voor het bewijs van (A1) noodzakelijke eigenschap luidt:

Is $\underline{z} \sim N_p(\mu, \sum_{(pp)})$, dan is de voorwaardelijke verdeling van $\underline{z}^{(2)}$ bij gegeven $\underline{z}^{(1)}$ gelijk aan:

$$f(\underline{z}^{(2)} | \underline{z}^{(1)}) = N_{p_2} \left[\mu^{(2)} + \sum^{21} \sum^{11^{-1}} (\underline{z}^{(1)} - \mu^{(1)}), \sum^{22 \cdot 1} \right] \quad (\text{A6})$$

waarin:

$$\sum_{(p_2 p_2)}^{22 \cdot 1} = \sum^{22} - \sum^{21} \sum^{11^{-1}} \sum^{12} \quad (\text{A7})$$

De uitdrukking (A7) welke de variantie van $\underline{z}^{(2)}$ bij gegeven $z^{(1)}$ voorstelt, is dus analoog aan (A5).

Het bewijs van (A1), namelijk:

$$\left| \sum_{(pp)} \right| = \prod_{i=1}^p \sigma_i^2 \quad |R| = \sigma_1^2 \sigma_{2.1}^2 \sigma_{3.12}^2 \cdots \sigma_{p.12\dots(p-1)}^2 \quad (A8)$$

zal worden gegeven met behulp van volledige inductie. De relatie (A8) is juist voor $p = 1$.

Immers dan geldt:

$$\sigma_1^2 = \sigma_1^2$$

Stel nu dat (A8) juist is voor p variabelen; dit betekent dat de volgende uitdrukking:

$$\left| \sum_{(pp)} \right| = \prod_{i=1}^p \sigma_i^2 \quad |R| = \sigma_1^2 \sigma_{2.1}^2 \sigma_{3.12}^2 \cdots \sigma_{p.12\dots(p-1)}^2 \quad (A9)$$

juist is.

Dan moet nu bewezen worden, dat de relatie (A8) geldt voor $(p+1)$ variabelen, dat wil zeggen:

$$\left| \sum_{(p+1 \ p+1)} \right| = \prod_{i=1}^{p+1} \sigma_i^2 \quad |R| = \sigma_1^2 \sigma_{2.1}^2 \sigma_{3.12}^2 \cdots \sigma_{(p+1).12\dots p}^2$$

Rekening houdend met (A9) betekent dit, dat moet gelden:

$$|\sum_{(p+1 \ p+1)}| = |\sum_{(pp)}| \cdot \sigma_{(p+1).12\dots p}^2 \quad (\text{A10})$$

Indien nu de $(p+1)$ variabelen $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_{p+1}$ worden gesplitst in 2 groepen; één van p variabelen en één van 1 variabele, met andere woorden:

$$\underline{z}^{(1)} = \begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \\ \vdots \\ \underline{z}_p \end{pmatrix} \quad \underline{z}^{(2)} = \underline{z}_{p+1}$$

dan betekent dit een partitionering van $\sum_{(p+1 \ p+1)}$ volgens:

$$\sum_{(p+1 \ p+1)} = \begin{pmatrix} \sum_{(pp)}^{11} & \sum_{(p1)}^{12} \\ \sum_{(1p)}^{21} & \sum_{(11)}^{12} \end{pmatrix}$$

waarbij:

$$\sum_{(11)}^{22} = \sigma_{p+1}^2$$

Gebruikmakend van (A4), is de determinant van $\sum_{(p+1 \ p+1)}$ te schrijven als:

$$\left| \sum_{(p+1 \ p+1)} \right| = \left| \sum_{(pp)}^{11} \right| \left| \sum_{(11)}^{22.1} \right| \quad (A11)$$

waarin, analoog aan (A5):

$$\sum_{(11)}^{22.1} = \sum^{22} - \sum^{21} \sum^{11^{-1}} \sum^{12}$$

Deze uitdrukking geeft echter, overeenkomstig (A7), ook de voorwaardelijke variantie weer van $\underline{z}^{(2)}$ bij gegeven $\underline{z}^{(1)}$.

In onze situatie betekent dat dus de voorwaardelijke variantie van de variabele \underline{z}_{p+1} bij gegeven z_1, z_2, \dots, z_p .

$\sum_{(11)}^{22.1}$ is dus een scalar, en wel gelijk aan $\sigma_{(p+1),123\dots p}^2$, waarmee

(A10), en daarmee ook (A1), is bewezen.

SUMMARY

The object of this study is to analyse the time dependent character of the components obtained with the aid of the method of the principal components, by considering their time dependency in relation to that of the original variables, which in the first instance, determine the model.

Within the model, formed by a component itself, the mutual time dependency of the individual variables may be studied at the same time.

This approach makes the not so successful cross spectral analysis for most economic time series, completely superfluous.

The starting point of this study is formed by the components to which, whether upon rotation or not, we have tried to give a significant interpretation.

Each component consists of a linear combination of the original variables $\underline{z}_1, \underline{z}_2, \dots, \underline{z}_p$.

Indicating the i -th component with \underline{f}_i , then:

$$\underline{f}_i = b_{i1} \underline{z}_1 + b_{i2} \underline{z}_2 + \dots + b_{ij} \underline{z}_j + \dots + b_{ip} \underline{z}_p \quad (1)$$

$$(i=1,2,\dots,p)$$

The coefficient b_{ij} of \underline{z}_j is indicating the measure of the importance of the variable \underline{z}_j in the component \underline{f}_i .

The above expression, which demonstrates in some way the mutual interrelationship between the variables, may be considered as a model to describe the actually non measurable component \underline{f}_i . This model is also part of the overall description of the structural dependency between the variables.

Expressing all the variables \underline{z}_j from (1) as a function of time, then the time dependent character of the corresponding component can be studied.

For that purpose the stochastic proces $\underline{z}_j(t)$, determined by the stochastic variable \underline{z}_j and the N observations taken at the moments $t=1,2,\dots,N$, is represented by a canonical expansion according to:

$$\begin{aligned}\underline{z}_j(t) &= \sum_{k=1}^n \underline{x}_{jk} \Psi_k(t) \\ &= \underline{x}_{j1} \Psi_1(t) + \underline{x}_{j2} \Psi_2(t) + \dots + \underline{x}_{jk} \Psi_k(t) + \dots + \underline{x}_{jn} \Psi_n(t) \\ &\quad (2) \\ &\quad (j=1,2,\dots,p)\end{aligned}$$

In the above expression the functions $\Psi_k(t)$, being only dependent upon time, are non random, while the coefficients \underline{x}_{jk} are assumed to be random.

From rewriting $\underline{z}_j(t)$ according to (2) in terms of the time dependent functions $\Psi_k(t)$, the similarity in train of thought appears which underlies the method of the canonical expansion of stochastic processes and that of the analysis of principal components.

The components analysis amounts to rewriting the variables \underline{z}_j in terms of the intrinsic variables or components $\underline{f}_1, \underline{f}_2, \dots, \underline{f}_p$. The expression (2) into which the variable \underline{z}_j is rewritten in terms of the functions $\Psi_1(t), \Psi_2(t), \dots, \Psi_n(t)$, to be considered as coordinates, essentially reflects the same principle.

The way of thinking with regard to the derivation of the optimal coordinate functions, the so -called Karhunen-Loève functions $\Phi_k(t)$ - to be shown in Chapter II- is absolutely analogous to that of the derivation of the components, as will be given in the second paragraph of Chapter I.

The derivation of the Karhunen-Loève coordinate functions may be considered as a "one dimensional components analysis" in which the

function $\Phi_k(t)$ is the analogon of the component \underline{f}_k .

Under -optimal- should be understood that a maximum explanation of the variables is obtained with a minimum of coordinate functions. The coordinate system of functions $\{\Phi_k(t)\}$, appears to be optimal, if these functions are the eigenfunction solutions of a homogeneous integral equation.

The kernel of this integral equation is formed by the correlation function of the corresponding variable \underline{z}_j .

In Chapter III, a method will be developed for the solution of this integral equation.

It appears that the properties of \underline{z}_j , such as e.g. its trend, periodicity and others may be inserted in a significant manner into the optimal solution $\Phi_1(t)$, $\Phi_2(t)$, ..., $\Phi_n(t)$.

Substitution of (2) in (1) gives a description of \underline{f}_i in terms of the time dependent properties of the variables \underline{z}_1 , \underline{z}_2 , ..., \underline{z}_p . The product b_{ij} with x_{jk} is the measure for the importance of the function $\Phi_k(t)$, belonging to the variable \underline{z}_j , in the component \underline{f}_i . These measures of importance may certainly be of interest when testing special views, whether subjective or not, with regard to the time sensitivity of the variables, which are involved in the econometric model.

LITERATUUR

Anderson, T.W.

The use of factor analysis in the statistical analysis of multiple time series.
Psychometrika - 1963, Vol. 28, No. 1, Blz. 1.

Baggeroer, A.B.

State variables and communication theory.
The M.I.T. Press: Research Monograph No. 61, 1970.

Bijnen, E.J.

Cluster - Analyse, overzicht en evaluatie van technieken.
Dissertatie, 1969.

Billingsley, P.

Ergodic theory and information.
Wiley, New York, 1965.

Bloxom, B.

Factorial rotation to simple structure and maximum similarity.
Psychometrika - 1968, Vol. 33, No. 2, Blz. 237.

Bosch, A.J.

Inleiding in de multivariate analyse I.
Technische Hogeschool Eindhoven, 1971.

Breipohl, A.M.

Probabilistic systems analysis;
an introduction to probabilistic models, decisions, and
applications of random processes.
Wiley, New York, 1970.

Browne, M.W.

A comparison of factor analytic techniques.
Psychometrika - 1968, Vol. 33, No. 3, Blz. 267.

Browne, M.W.

A note on lower bounds for the number of common factors.
Psychometrika - 1968, Vol. 33, No. 2, Blz. 233.

Browne, M.W.

Fitting the factor analysis model.
Psychometrika - 1969, Vol. 34, No. 3, Blz. 375.

Browne, M.W. Kristof, W.

On the oblique rotation of a factor matrix to a specified pattern.
Psychometrika - 1969, Vol. 34, No. 2, Blz. 237.

- Corballis, M.C. Traub, R.E.
Longitudinal factor analysis
Psychometrika - 1970, Vol. 35, No. 1, Blz. 79.
- Ferrari, Th.J. Mol, J.
Factor analysis of causal models.
Neth. J. Agric. Sci. - 1967, Vol. 15, Blz. 48.
- Fishman, G.S.
Spectral methods in econometrics.
Harvard University Press, Cambridge, Massachusetts, 1969.
- van de Geer, J.P.
Inleiding in de multivariate analyse
van Loghum Slaterus, Arnhem, 1967.
- Gollob, H.F.
A statistical model which combines features of factor
analytic and analysis of variance techniques.
Psychometrika - 1968, Vol. 33, No. 1, Blz. 73.
- Granger, W.J. Hatanaka, M.
Spectral analysis of economic time series.
Princeton University Press, Princeton, 1964.
- Guertin, W.H. Bailey, J.P.
Introduction to modern factor analysis.
Ann Arbor, Michigan, 1970.
- Harman, H.H. Jones, W.H.
Factor analysis by minimizing residuals (MINRES).
Psychometrika - 1966, Vol. 31, No. 3.
- Harris, Ch.W.
Some Rao-Guttman relationships.
Psychometrika - 1962, Vol. 27, No. 3, Blz. 247.
- Harris, Ch.W.
Problems in measuring change.
The University of Wisconsin Press, Madison, 1963.
- Harris, Ch.W.
On factors and factor scores.
Psychometrika - 1967, Vol. 32, No. 4.
- Heermann.
The algebra of factorial indeterminacy.
Psychometrika - 1966, Vol. 31, No. 4.

Holtzman, W.H.

Statistical methods for the study of change in the single case.

zie: Harris, Ch.W.

Problems in measuring change; Blz. 199.

Hope, K.

Methods of multivariate analysis.

University of London Press Ltd., 1968.

Horst, P.

Factor analysis of data matrices.

Holt, Rinehart and Winston, Inc. 1965.

Horst, P.

Multivariate models for evaluating change.

zie: Harris, Ch.W.

Problems in measuring change; Blz. 104.

Jenkins, G.M.

General considerations in the analysis of spectra.

Technometrics - 1961, Vol. 3, No. 2.

Jöreskog, K.G.

Some contributions to maximum likelihood factor analysis.

Psychometrika 1967, Vol. 32, No. 4.

Krzanowski, W.J.

The algebraic basis of classical multivariate methods.

The Statistician - 1971, Vol. 20, No. 4, Blz. 51.

Kullback, S.

Information theory and statistics.

Wiley, New York, 1959.

Lawley, D.N.

Maxwell, A.E.

Factor analysis as a statistical method.

Butterworths, London, 1963.

Linfoot, E.H.

An informational measure of correlation.

Information and Control - 1957, No. 1, Blz. 85 - 89.

Lord, F.M.

Elementary models for measuring change.

zie: Harris, Ch.W.

Problems in measuring change; Blz. 21.

Malinvaud, E.

Statistical methods of econometrics.
North-Holland Publishing Company - Amsterdam, 1966.

Mc Donald, R.M.

A general approach to nonlinear factor analysis.
Psychometrika - 1962, Vol. 27, No. 4, Blz. 397.

Meyers, S.L.

A factor analysis approach to studying the structure of the
stock market.
Dissertatie, 1970.

Mol, J. Ferrari, Th.J.

Factor analysis of causal models.
Neth. Journal Agric. Sci. 1967, 15. Blz. 38 - 49.

Mol, J.

Confrontatie van factoranalyse en regressie analyse.
Instituut voor economisch onderzoek van de Rijks Universiteit
te Groningen, 1969.

Pugachev, V.S.

Theory of random functions and its application to control
problems.
Pergamon Press, Oxford, 1965.

Rényi.

Warscheinlichkeitsrechnung,
mit einem Anhang über Informationstheorie.
VEB Deutscher Verlag der Wissenschaften, Berlin, 1962.

Ross, J.

Informational coverage and correlational analysis.
Psychometrika - 1962, Vol. 27, No. 3, Blz. 297.

Schilderlinck, J.H.F.

Factor analysis applied to developed and developing countries.
Rotterdam University Press,
Wolters-Noordhoff Publishing, Groningen, 1970.

Sherin, R.J.

A matrix formulation of Kaiser's Varimax criterion.
Psychometrika - 1966, Vol. 31, No. 4.

Solodovnikov, V.V.

Introduction to the statistical dynamics of automatic control
systems.
Dover Publications, Inc. 1960.

Stobberingh, R.

The derivation of the optimal Karhunen-Loève coordinate functions.

Tilburg Institute of Economics, Research Memorandum 24, 1971

Thorndike, R.M.

Method of extraction, type of data, and adequacy of solutions in factor analysis.

Dissertatie, 1970.

Thurstone, L.L.

Multiple-factor analysis.

The University of Chicago Press, 1965.

Tou, J.T. (ed.)

Advances in information systems science.

Plenum Press, New-York, 1969 - 1970.

Überla, K.

Faktorenanalyse.

Springer-Verlag, Berlin, 1968.

Watanabe, S.

Information theoretical analysis of multivariate correlation.

IBM Journal of Research

and Development, 1960, 4, Blz. 66 - 82.

Watanabe, S.

Feature compression; The role of "features" in pattern recognition.

zie: Tou, J.T.

Advances in information systems science, 1969 - 1970. Hoofdstuk II.

gr
u qre

supreme

na en

4.

a

as

ut

Δ ()

le, al

UCOMMIT

COMMIT

COMMIT

a rep

an

te de

an rep

etacr

an or